
# Big Data en el sector sanitari

**Miquel Morey Riera**
Grau Enginyeria Informàtica
TFG - Business Intelligence

**Humberto Andrés Sanz**
**Atanasi Daradoumis Haralabus**

15/06/2016

## FITXA DEL TREBALL FINAL

| | |
|---|---|
| **Títol del treball:** | *Big Data en el sector sanitari* |
| **Nom de l'autor:** | *Miquel Morey Riera* |
| **Nom del consultor/a:** | *Humberto Andrés Sanz* |
| **Nom del PRA:** | *Atansi Daradoumis Haralabus* |
| **Data de lliurament (mm/aaaa):** | *06/2016* |
| **Titulació o programa:** | *Grau Enginyeria Informàtica* |
| **Àrea del Treball Final:** | *Business Intelligence* |
| **Idioma del treball:** | *Català* |
| **Paraules clau** | *BigData Sanitat Cloud* |

**Resum del Treball**

La finalitat del treball és analitzar com Big Data Analytics pot aportar valor al sector sanitari repercutint en estalvis econòmics i millorant l'atenció sanitària. En el marc econòmic actual de crisis generalitzada i pressuposts limitats, també es requereix a les àrees de TI l'aportació de valor amb el mínim de recursos. Les solucions SaaS de Big Data Analytics que comencen a aparèixer poden ser una alternativa econòmica per obtenir valor de Big Data. Així mateix poden aportar avantatges significatius en l'anàlisi estadística.

S'introdueix el concepte i les tecnologies Big Data en amplitud i es comparen els seus avantatges i limitacions amb les tecnologies tradicionals. També s'aborden els coneixements necessaris per el tractament i anàlisi de Big Data juntament amb els aspectes rellevants i aportacions de l'analítica d'aquestes dades al sector sanitari. Això permet introduir un model d'arquitectura genèrica Big Data Analitycs per aquest sector concret.

Les barreres existents en el sector sanitari, principalment el tractament de dades de caràcter personal, dificulten les solucions Big Data Analytics en núvols públics. Els models de presentació de serveis de Cloud Computing ofereixen una flexibilitat i escalabilitat excepcionals i, més concretament els models híbrids, permeten mitigar el risc i reduir costos.

Actualment, el sector sanitari té dificultats per explorar eficientment les seves dades però ja existeixen algunes aplicacions en modalitat SaaS de varies organitzacions que estan impulsant l'evolució de l'analítica Big Data en Sanitat. Pot ser, el futur està en combinar adequadament Big Data Analytics i Cloud Coputing per aconseguir un ecosistema millor.

**Abstract**

The aim of this work is to analyze how Big Data Analytics can contribute to give an added value to healthcare system, having repercussions in terms of economic savings and improving healthcare as a whole. The current economic framework, with its generalized crisis and limited budgets, is also demanding the IT departments to get more value but with fewer resources. SaaS solutions of Big Data Analytics that are emerging may be a cheap alternative to get this value from Big Data. Likewise, they can give significant advantages in statistic analysis.

Big Data concept and technologies will be widely introduced and its advantages and limitations will be compared to traditional technologies. An approach of the needed knowledge to deal and analyze Big Data will also be made, along with other relevant aspects and contributions of this analysis to healthcare system. All of this will allow us to introduce a generic model of architecture of Big Data Analytics in this particular sector.

Existing barriers in healthcare sector, above all things personal data processing, hinder carrying Big Data Analytics to the Cloud. Cloud Computing services offer a highly flexible and scalable solutions and, in case of hybrid models, allow to mitigate risks and reduce expenses.

Nowadays, healthcare system is in big trouble to efficiently explore its data, but there are already some SaaS based applications in several organizations that are boosting Big Data Analytics in healthcare. The future may be in properly combining Big Data Analytics and Cloud Computing to achieve an improved ecosystem.

# Índex

## Llista de figures

# 1. Introducció

La temàtica d'aquest Treball Final de Grau (TFG) versarà sobre les solucions Business Intelligence (BI) en el sector sanitari, concretament amb la utilització de la tecnologia Big Data Analytics.

Els centres sanitaris i els pacients generen i acumulen grans quantitats de dades en molts diferents formats, ja sigui en paper o en versió digital, però que no són utilitzades per els sistemes BI tradicionals per la dificultat i/o impossibilitat material de tractar-les de forma efectiva.

Big Data és un relativament nou sistema de suport a la presa de decisions basades en fets que aprofita el gran volum de dades que es generen actualment. Els tres atributs que el caracteritzen: Volum escalable d'emmagatzemament, Velocitat en incorporar i analitzar dades i Varietat de formats de dades (estructurades, semiestructurades i gens estructurades) que pot tractar; ofereixen a les organitzacions que l'utilitzen correctament noves fonts de coneixement que no proporcionen els sistemes BI tradicionals.

D'aquesta manera, el sector sanitari és un sectors on Big Data pot tenir un gran impacte en l'actualitat ja que possibilita l'anàlisi de totes aquestes dades de manera més àmplia i eficient i aportar importants avantatges[1]. Per una banda, repercutint en estalvis econòmics en quant a:

- Millora de la coordinació en l'atenció al pacient
- Detecció de fraus i abusos
- Detecció d'ineficiències administratives i clíniques.

Per altra banda, millorant l'atenció sanitària:

- Suport a la investigació
- Transformació de les dades en informació
- Suport al auto cuidat
- Suport als proveïdors de cuidats mèdics
- Augment de coneixement i conscienciació de l'estat de salut
- Agrupament de les dades per expandir l'ecosistema.

Concretament, en la primera part del treball, s'aprofundirà en el concepte de Big Data, s'avaluaran les necessitats de noves tecnologies per tractar de manera eficient aquests tipus de dades.

També s'analitzarà la plataforma Big Data per tal de conèixer la infraestructura subjacent, tant els recursos físics com lògics, com el seu funcionament. Seguint amb la mateixa anàlisi, també caldrà conèixer els mecanismes per incorporar les dades a aquesta plataforma, els sistemes d'emmagatzemament i anàlisi que s'utilitzen per poder visualitzar el coneixement i, en conseqüència, treure valor d'aquestes dades.

Juntament amb aquesta primera part, es revisaran les diferents alternatives d'allotjament que es poden utilitzar per implementar els serveis Big Data Analytics tant "*on premise*" com en els diferents modalitats de "*Cloud Computing*". Es posarà especial èmfasi en aquest darrer, concretament amb proveïdors que proporcionin eines analítiques avançades en modalitat *Software as a Service* (SaaS).

El TFG no es limitarà solament a l'anàlisi descriptiva de la plataforma Big Data, sinó que avançarà en els criteris que cal que es considerin en el sector de la sanitat. Per això, en la segona part del treball s'analitzaran els requeriments per una arquitectura Big Data per el sector sanitari i possibles aplicacions de Big Data Analytics per aquest sector concret. A més, es sospesaran aspectes el cost d'implantació de les solucions i/o l'estalvi en optar per l'aplicació al núvol confrontada amb una aplicació d'aquestes característiques *on premise*.

## 1.1. Context i justificació del Treball

L'actual entorn econòmic turbulent que acusa els efectes de molts anys consecutius de crisi generalitzada, propicia que les empreses realitzin un esforç suplementari en la gestió i optimització dels costos. Centrant-nos en l'àrea de les tecnologies de la informació (TI) tampoc no es pot escapar d'aquesta dinàmica restrictiva. Així, les empreses imposen control pressupostari sever sobre les alternatives de gestió de les dades i les solucions Business Intelligence (BI) implantades o noves solucions *Analytics*. Com més va, el paradigma del núvol i les solucions *Software as a Service* (SaaS) constitueixen opcions vàlides a l'hora d'assolir els objectius de disciplina pressupostària dels departaments TI i de l'empresa considerada en la seva totalitat. Darrerament apareixen proveïdors de solucions BI consolidats i noves *startups* que implementen solucions Big Data Analytics en règim SaaS i els CIOs de les companyies tenen de diferents alternatives a considerar[2]. El propòsit últim d'aquesta memòria no va més enllà de contribuir a aportar elements i criteris que permetin avançar en cada cas de selecció que es plantegi.

Les implantacions locals de solucions BI, i més encara les de Big Data Analytics, acostumen a comportar requisits relativament elevats pel que fa a la infraestructura necessària que s'acompanyen d'una inversió financera quantiosa. Per contra, en general, les solucions BI al núvol ofereixen avantatges, operatius i financers, que l'entorn corporatiu aprecia i que, amb freqüència creixent, contribueixen a la migració des de les implantacions locals envers implantacions de paradigma SaaS. Remarquem les més importants:

- Rapidesa en la implantació i el desplegament de les solucions BI: la disponibilitat immediata de la solució escollida, sense cap mena de dependència dels llargs períodes d'espera que normalment s'associen al subministrament de la infraestructura i el desplegament del programari, redueix dràsticament la finestra temporal d'implantació d'una aplicació BI.

- Escalabilitat/Elasticitat: derivada de l'aprofitament del poder de computació massiu disponible a la Web, amb la possibilitat d'expandir o

reduir capacitat d'acord amb les necessitats empresarials de cada moment.

- Enfocament en les competències crítiques de l'empresa: la implantació de solucions al núvol permet externalitzar una bona part de la gestió BI a professionals d'aquest àmbit i allibera recursos del departament TI que poden focalitzar la seva activitat en competències "core".

- Reducció del cost total de propietat de la solució: des d'una òptica merament financera una fracció de la inversió fixa (*capital expenditure*) de l'empresa es transforma en despesa operativa (*operational expenditure*) amb els avantatges fiscals que se'n poden derivar. Addicionalment, la política de preus dels proveïdors de les solucions i/o els models "*pay-per-use*" acostumen a resultar prou competitius.

- Àmplia disponibilitat: que permet servir als usuaris mòbils i remots. En general, l'accés a la l'aplicació basat en el navegador, permet el control absolut des de la plataforma al núvol fins a la gestió de les bases de dades, des del emmagatzemament fins a les eines analítiques.

## 1.2. Objectius del Treball

L'objectiu bàsic d'aquest treball consisteix essencialment en la realització d'un anàlisi d'enfocament general de la plataformes Big Data Analytics. L'abast es limitarà a algunes solucions que justifiquin les aportacions i potencial en l'entorn sanitari, o que presentin algun tret diferencial que les faci particularment interessants. S'analitzaran les capacitats i els principals avantatges que poden aportar les plataformes Big Data Analytics a l'hora de facilitar les decisions en aquest sector.

El treball no es limitarà, però, a la mera revisió de característiques tècniques i a la identificació d'avantatges i febleses la plataforma. També es realitzarà una aproximació als criteris corporatius sanitaris a l'hora de decidir quina solució particular pot resultar adient per a la implantació de determinats projectes Big Data.

Es voldrà emfasitzar en la vessant del retorn sobre la inversió que es pot assolir amb la implantació d'una solució en núvol de Big Data Analytics. Retorn estimable tant en termes financers (increment d'ingressos, disminució de costos o tots dos objectius alhora) com en termes no estrictament financers (millores en el nivell de satisfacció dels pacients, millores en l'eficiència del personal sanitari, millores en el grau de confiança del procés de presa de decisions, ...)

## 1.3. Enfocament i mètode seguit

La memòria que ens ocupa pretén apartar-se d'un hipotètic enfocament merament acadèmic i opta per centrar l'atenció en l'òptica de caire empresarial que es preocupa per l'optimització dels recursos (escassos per definició) existents al si de les corporacions i la selecció de les millors alternatives que els esmentats recursos permeten utilitzar. Resumidament, aquesta memòria considerarà de forma prioritària els aspectes més pràctics (restriccions tècniques, consideracions de cost/estalvi econòmic, ...) que intervenen en el procés de selecció d'una determinada solució Big Data Analitycs per part del sector sanitari.

El ventall de proveïdors de solucions Big Data Analytics amb modalitat SaaS és emergent. L'autor d'aquest treball opta preferentment per l'accés a fonts primàries: accés a la documentació / especificació tècnica i oferta de cada solució a l'hora d'avaluar la funcionalitat i les diferents dimensions tècniques de les solucions revisades. Tanmateix, s'acudirà a fonts secundàries (llibres especificats a la bibliografia, publicacions online diverses, ...) per tal d'introduir els conceptes bàsics relatius al paradigma del núvol i les aplicacions SaaS i els relatius a l'àmbit específic de Big Data Analytics.

La problemàtica que planteja la selecció d'una solució Big Data Analytics presenta un nivell de complexitat relativament elevat i depèn, també, d'un nombre elevat de dimensions que cal considerar en cada projecte concret. En qualsevol cas i a partir de la informació subministrada per les fonts primàries, l'autor s'esforçarà en identificar i objectivar diferents criteris que hauran de facilitar el procés de selecció d'una solució determinada segons les característiques i necessitats de cada area sanitària que estudiï la implantació i desplegament d'una solució Big Data Analytics. L'aplicació dels criteris als projectes Big Data permetrà obtenir les principals conclusions que es derivaran al capítol final de la memòria.

## 1.4. Planificació del Treball

S'ha procedit a temporalitzar les tasques requerides per a l'elaboració d'aquest Treball de Fi de Grau, tot respectant les dates clau de lliurament de les diferents activitats d'avaluació continuada i de la memòria final. Seguidament es copia la temporalització i diagrama de Gannt realitzats amb l'ajut de l'aplicatiu ProjectLibre:

| | | Nombre | Duracion | Inicio | Terminado |
|---|---|---|---|---|---|
| 1 | | **TFG/Big Data en entorns sanitaris** | **80 days** | **25/02/16 8:00** | **15/06/16 17:00** |
| 2 | | **Elaboració PAC1** | **15 days** | **25/02/16 8:00** | **16/03/16 17:00** |
| 3 | | Definició preliminar de temàtica i abast | 3 days | 25/02/16 8:00 | 29/02/16 17:00 |
| 4 | | Recerca bibliogràfica, fots primàries i secundàries | 7 days | 1/03/16 8:00 | 9/03/16 17:00 |
| 5 | | Aprovació idea i treball de redacció | 4 days | 10/03/16 8:00 | 15/03/16 17:00 |
| 6 | | Lliurement de pla de Treball - PAC1 | 1 day | 16/03/16 8:00 | 16/03/16 17:00 |
| 7 | | **Elaboració PAC2** | **25 days** | **17/03/16 8:00** | **20/04/16 17:00** |
| 8 | | Breu revisió conceptes Basics Cloud Computing | 1 day | 17/03/16 8:00 | 17/03/16 17:00 |
| 9 | | Cerca de sistemes Big Data | 2 days | 18/03/16 8:00 | 21/03/16 17:00 |
| 10 | | **Estudi tecnologies Big Data** | **7 days** | **22/03/16 8:00** | **30/03/16 17:00** |
| 11 | | Conceptes i programaris | 2 days | 22/03/16 8:00 | 23/03/16 17:00 |
| 12 | | Requeriments tecnològics | 2 days | 25/03/16 8:00 | 28/03/16 17:00 |
| 13 | | Eines de captura, anàlisi i visualització | 2 days | 29/03/16 8:00 | 30/03/16 17:00 |
| 14 | | Integració feedback PAC1 (Modificacions adients) | 1 day | 31/03/16 8:00 | 31/03/16 17:00 |
| 15 | | Anàlisi dels components i capacitats Big Data | 7 days | 1/04/16 8:00 | 11/04/16 17:00 |
| 16 | | Redacció PAC2: Tecnologies Big Data i capacitats | 6 days | 11/04/16 17:00 | 19/04/16 17:00 |
| 17 | | Lliuranment PAC2 | 1 day | 20/04/16 8:00 | 20/04/16 17:00 |
| 18 | | **Elaboració PAC3** | **25 days** | **21/04/16 8:00** | **25/05/16 17:00** |
| 19 | | **Big Data en Sanitat** | **7 days** | **21/04/16 8:00** | **29/04/16 17:00** |
| 20 | | Anàlisi del sector sanitari | 3 days | 21/04/16 8:00 | 25/04/16 17:00 |
| 21 | | Integració de la plataforma Big Data en el sector | 2 days | 26/04/16 8:00 | 27/04/16 17:00 |
| 22 | | Anàlisi d'avantatges, riscos i barreres | 1 day | 28/04/16 8:00 | 28/04/16 17:00 |
| 23 | | Analisi de seguretat (Protecccció de dades personals) | 1 day | 29/04/16 8:00 | 29/04/16 17:00 |
| 24 | | Integració feedback PAC2 (Modificacions adients) | 2 days | 30/04/16 8:00 | 3/05/16 17:00 |
| 25 | | **Big Data en Cloud** | **4 days** | **4/05/16 8:00** | **9/05/16 17:00** |
| 26 | | Consideracions d'infraestructura | 2 days | 4/05/16 8:00 | 5/05/16 17:00 |
| 27 | | Anàlisi d'avantatges, inconvenients | 2 days | 6/05/16 8:00 | 9/05/16 17:00 |
| 28 | | Aplicacions SaaS Big Data en sanitat | 3 days | 10/05/16 8:00 | 12/05/16 17:00 |
| 29 | | Redacció PAC3: Big Data en Sanitat, Cloud i aplicacions | 6 days | 13/05/16 8:00 | 20/05/16 17:00 |
| 30 | | Conclusions del treball | 2 days | 23/05/16 8:00 | 24/05/16 17:00 |
| 31 | | Lliurament PAC3 | 1 day | 25/05/16 8:00 | 25/05/16 17:00 |
| 32 | | **Redacció Final TFG** | **15 days** | **26/05/16 8:00** | **15/06/16 17:00** |
| 33 | | Agregació PCS1/PAC2/PAC3 i síntesi, redacció | 3 days | 26/05/16 8:00 | 30/05/16 17:00 |
| 34 | | Integració feedback PAC3/modificacions adients | 1 day | 1/06/16 8:00 | 1/06/16 17:00 |
| 35 | | Revisió, lectura i refinament de memòria completa | 3 days | 2/06/16 8:00 | 6/06/16 17:00 |
| 36 | | Preparació i gravació vídeo presentació | 3 days | 7/06/16 8:00 | 9/06/16 17:00 |
| 37 | | Elaboració detallada abstract en Català i Anglès | 2 days | 10/06/16 8:00 | 13/06/16 17:00 |
| 38 | | Lliurament de la presentació en vídeo | 1 day | 13/06/16 8:00 | 13/06/16 17:00 |
| 39 | | Informe d'autoavaluació de competències transversals | 1 day | 14/06/16 8:00 | 14/06/16 17:00 |

**Figura 1. Temporització**

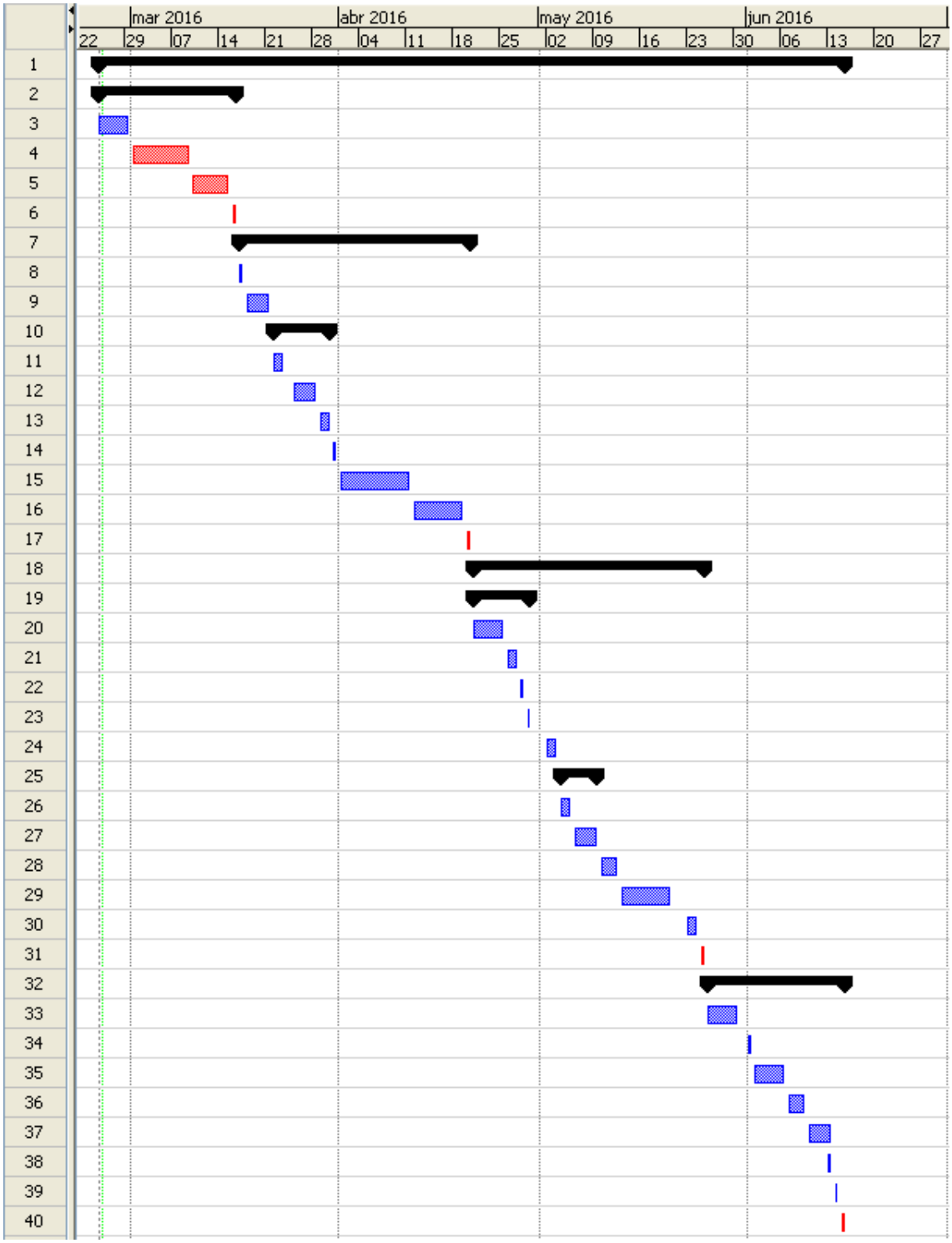


**Figura 2. Diagrama de Gannt**

## 1.5. Sumari de productes obtinguts

La naturalesa d'aquest treball no comporta la codificació de cap tipus de programari. Conseqüentment, els lliurables principals del treball consistiran en els tres lliuraments parcials que constitueixen les Proves d'Avaluació Continuada (PAC) numerades 1, 2 i 3 i, per descomptat, el lliurament de la memòria final del Treball de Fi de Grau, degudament acompanyat per una breu presentació de defensa del projecte en format audiovisual.

## 1.6. Breu descripció dels altres capítols de la memòria

En el capítol 2 "*Big Data. Raó i definició*", a grans trets, es defineix el concepte de Big Data. Per entrar en detall, en primer lloc, s'analitzen els orígens d'aquestes dades, es a dir, qui les genera, com i amb quina quantitat. També s'entra en el detall de com són aquestes dades fent referència a la seva estructura. S'aprofundeix en les necessitats de tecnologies diferents a les tradicionals per el tractament de les dades, ja que aquestes darreres tenen unes quantes limitacions que no permeten el tractament adequat de les mateixes. Així mateix, en el marc de la BI (intel·ligència de negoci), es justifiquen els avantatges de les eines analítiques de Big Data en front de les BI tradicionals per poder extreure informació útil a partir d'aquestes dades. Finalment, es defineixen els atributs que caracteritzen Big Data i s'introdueix la seva complexitat. D'aquesta manera s'aconsegueix tenir una base més sòlida per entendre millor les dades que generen els sistemes d'informació sanitaris i la possibilitat d'analitzar-les juntament amb altres fonts de dades disponibles per obtenir coneixement i optimització de recursos.

La "*Ciència de les dades*", títol del capítol 3, presenta la complexitat inherent al tractament de Big Data. La gran quantitat de dades, la velocitat amb que es generen i ,fins i tot s'analitzen, juntament amb la seva varietat, requereix d'amplis coneixements en certes àrees de la Intel·ligència Artificial, Estadística i Mineria de dades. Així mateix, també són necessaris coneixements sobre fonts de dades i les seves taxonomies, tècniques de processament massiu de dades, tècniques de tractament de dades, repositoris especialitzats d'emmagatzematge de Big Data, i tècniques i metodologies d'anàlisi i de visualització de dades que s'introdueixen a través dels diferents blocs d'aquest capítol. Totes aquestes tècniques s'implementaran d'una manera o altra en funció del problema concret que es pretén resoldre.

En el capítol 4 "*Tecnologies Big Data*" es presenten plataformes, components i eines que implementen les tècniques descrites en el capítol anterior. En aquest aspecte, Hadoop és la plataforma que ha popularitzat Big Data. Es descriuen els seus principals components, així com el conjunt d'eines vinculades que permeten la seva orquestració per el tractament de dades. L'evolució de Big Data fa que constantment apareguin noves eines i evolucions de la plataforma, per això es destaquen els motors de processament avançats *Apache Spark* i *Apache Storm* que permeten el processament i anàlisi de dades en temps real o quasi real. Addicionalment es fa referència a alguns dels principals proveïdors

que ofereixen les seves distribucions de sistemes Big Data Analitycs basades en aquesta plataforma.

Per emmarcar Big Data en el sector sanitari, en el capítol 5 es fa una anàlisi superficial d'aquest sector juntament amb el potencial i la problemàtica que se'n deriva de l'ús de Big Data en Sanitat. Un cop analitzades les fonts de dades que té disponibles i els resultats que se'n espera obtenir, es presenta un model de plataforma Big Data que integra tots els components per fer front al tractament del volum la varietat i la velocitat de les dades amb la finalitat d'obtenir resultats en aquest sector. Tanmateix, obtenir informació útil no tan sols depèn de les tecnologies implementades sinó també de la quantitat/qualitat de dades disponibles i de les capacitats humanes d'analitzar-les.

En l'objectiu de optimitzar costos, en el capítol 6 s'analitza el Cloud Computing com a facilitador de Big Data Analytics. Així, s'entra en el detall de com moltes organitzacions ja utilitzen alguns dels seus models i, en concret, com el model de núvol híbrid pot ser una combinació apropiada per el desplegament de solucions de Big Data Analytics. Amb aquesta combinació apareix el concepte emergent Big Data as a Service (BDaaS) que es presenta com un model de prestació de serveis que, al igual que Cloud Computing, permet utilitzar Big Data Analytics en les seves diferents modalitats. Finalment, es fa un resum de les avantatges d'utilitzar aquest model i com el sector sanitari es un dels que se'n pot beneficiar.

En el darrer capítol, Aplicacions "*BDaaS en Sanitat*" es presenten varies aplicacions que actualment són referents en l'àmbit sanitari en l'ús de Big Data Analytics i que ofereixen els seus serveis a través del núvol. Aquestes aplicacions s'agrupen en sis grups que es corresponen a les maneres on Big Data esta ajudant a l'assistència sanitària.

# 2. Big Data. Raó i definició

El concepte Big Data, que es pot traduir literalment com a 'dades massives', apareix per primera vegada en l'àmbit de les ciències d'astronomia i la genètica, tot i que la seva popularització apareix a mitjans de la passada dècada en l'àmbit de les TIC amb l'augment dels dispositius amb connexió a Internet, juntament amb l'auge de les xarxes socials. Moltes d'aquestes dades són obertes i accessibles per qualsevol que les vulgui explotar.

En aquest capítol s'aprofundirà en els orígens de Big Data i en el atributs que el caracteritzen. Es justificarà la necessitat de noves tecnologies per tractar aquestes dades davant les limitacions dels sistemes tradicionals i com Business Intelligence (BI) pot traure profit a l'hora de prendre decisions. Finalment, es descriu de manera més o menys formal el concepte Big Data.

## 2.1. D'on vénen les dades?

Els essers humans estem generant i emmagatzemant informació constantment i cada vegada en quantitats superiors. S'estima que la xifra es duplica cada dos anys segons l'estudi *EMC Digital Unviverse*[3] de 2014, també preveu que la quantitat de dades es multiplicarà per 10 entre 2013 i 2020 passant del 4,4 a 44 bilions de gigabytes, respectivament. Segons un altra informe, *Ericsson Mobility Report*[4] (agost 2014), el tràfic de xarxa augmenta vertiginosament, ja que va créixer un 60% entre el segon trimestre del 2013 i el del 2014.

La contribució a la acumulació de dades massives es pot trobar en diverses industries[5]. Les companyies generen i mantenen grans quantitats de dades amb informació de clients, proveïdors, operacions, etc. allotjades en sistemes ERP, CRM, SCM, KM, comerç electrònic i altres solucions sectorials necessàries per la seva activitat. (per exemple, basta pensar com deuen ser les bases de dades d'Amazon, Google, Twitter o Facebook). El mateix succeeix en el sector públic, en molts països s'administren enormes bases de dades que contenen dades de cens de població, registres mèdics, imposts, etc.

A totes aquestes dades també es poden afegir les transaccions financeres realitzades en línia per dispositius mòbils, comentaris a xarxes socials, ubicació geogràfica mitjançant GPS, *wereables*, en definitiva, totes aquelles activitats realitzades mitjançant un dispositiu mòbil. Cisco en el seu estudi *VNI Mobile Forecast*[6] projecta que per l'any 2020 hi haurà 5.500 milions d'usuaris mòbils que generaran un tràfic de 367 exabytes per any –en comparació als 44 exabytes en 2015.

Per altra banda, la comunicació màquina a màquina (M2M machine-to-machine) també contribueix a l'enorme creixement de les dades mitjançant sensors connectats a la xarxa que estan incrustats en diversos tipus de dispositius com telèfons mòbils, mesuradors intel•ligents d'energia, automòbils y maquinaria industrial, etc. que generen i comuniquen grans quantitats de dades. Segons indica Gartner[7] en el seu estudi sobre el Internet de les Coses

(IoT), hi haurà uns 4.900 milions d'aquests sensors en 2015 i preveu uns 25 mil milions per 2020.

Moltes empreses i organitzacions que acumulen dades massives han trobat una nova línia de negoci en el tractament i venta d'aquestes dades, com (operadores telefòniques, entitats bancàries o elèctriques, entre altres). Altres organitzacions i administracions ofereixen les dades de manera oberta i gratuïta que es coneixen com - open data -.

D'aquesta manera, es pot accedir cada vegada a més fonts de dades i amb una major quantitat de dades que es recullen, comparteixen i s'analitzen cada dia, de diversa varietat i de multitud de nínxols diferents que es materialitzen en centenars (si no milers) de fonts de dades disponibles[8], a punt per ser utilitzades i analitzades per qualsevol que estigui disposat a buscar-les.

## 2.2. Estructura de les dades

La gran quantitat i varietat de dades generades per les diferents fonts d'informació esmentades a l'apartat anterior tenen diferents estructures. Aquestes dades, a grans trets, es poden classificar en tres tipus: estructurades, no estructurades i semiestructurades

- Les dades estructurades són les dades que tenen ben definits la seva longitud i el seu format, com les dates, els números o les cadenes de caràcters. Aquestes s'emmagatzemen en taules. Un exemple són les bases de dades relacionals i els fulls de càlcul.

- Les dades no estructurades són dades en el format tal com van ser recol·lectades, no tenen un format específic. No es poden emmagatzemar dins d'una taula ja que no es pot desgranar la seva informació a tipus bàsics de dades. Alguns exemples són els PDF, documents multimèdia, correus electrònics o documents de text.

- Les dades semiestructurades són les que no es limiten a camps determinats, però que contenen marcadors per separar els diferents elements. És una informació poc regular com per ser gestionada d'una forma estàndard. Aquestes dades posseeixen les seves pròpies metadades o etiquetes que descriuen els objectes i les relacions entre elles. Un exemple són els fitxers HTML o els XML.

## 2.3. Limitacions del processament de dades tradicional

En general, les organitzacions de forma tradicional venen utilitzant bases de dades relacionals (RDB) per organitzar, emmagatzemar i processar les seves dades. Però les RDB resulten ser massa lentes i costoses per a les necessitats dels sistemes Big Data, principalment per dos motius. En primer lloc, poca escalabilitat en quant a l'emmagatzemament de dades i en segon lloc, limitacions per el tractament de dades no estructurades. S'explica amb més detall a continuació.

- Les RDB tenen unes capacitats d'emmagatzematge limitades a màxims de creixement de pocs gigabytes diaris. En el cas d'incloure orígens de dades massives, es pot augmentar considerablement el volum d'informació i sobrepassar aquest màxim de creixement, el que podria afectar considerablement al rendiment del sistema. Això és degut a que les RDB no es poden distribuir de forma senzilla sobre diverses màquines ja que tenen limitacions d'escalabilitat i per tant, no són adequades per sistema Big Data.

  El fonament teòric l'ofereix el teorema CAP o teorema de *Brewer*, que diu que en sistemes distribuïts no és possible garantir alhora:
  - La Consistència (en anglès "*Consistency*") fent referència a que es rep la mateixa informació independentment del node que processi la petició.
  - La Disponibilitat (en anglès "*Availability*") es refereix a que tots els clients puguin llegir i escriure encara que algun node falli.
  - La Tolerància a les Particions (en anglès "*Partition tolerance*") vol dir que el sistema funciona encara que falli una partició.

  Les bases de dades relacionals compleixen bé les dues primeres, però les bases de dades del Big Data necessiten complir prioritàriament la Tolerància a les particions.

- Les RDB tenen limitacions per processar dades no estructurades. En general permeten emmagatzemar textos, documents i arxius multimèdia, però no disposen de funcionalitats addicionals per processar el seu contingut de manera eficient. Per processar o visualitzar la informació no estructurada emmagatzemada en aquest tipus de bases de dades solen requerir aplicacions de tercers, o extensions de la base de dades.

  Per el cas de tractament de textos, algunes RDB incorporen capacitats documentals que incorporen funcions de recerca en textos, documents i altres funcions de gestió documental i multimèdia, que permeten extreure metadades dels fitxers emmagatzemats, Tot i així, el seu tractament és limitat i ofereix una funcionalitat limitada.

## 2.4. Necessitat de noves tecnologies

L'evolució de la tecnologia va sorgir de grans empreses d'Internet, com Google, Yahoo i Amazon. Aquestes acumulaven i tractaven grans quantitats de dades i els sistemes de processament tradicionals no permetien tractar-les de manera eficient. Per fer front al tractament de les noves fots de dades es requerien noves eines i tecnologies considerant els següents aspectes:

- La gran quantitat de dades fan inviable el seu processament en un únic ordinador, per gran i potent que sigui. Calen sistemes de processament distribuït per integrar diferents ordinadors que treballin amb les dades de manera paral·lela per tal de processar més dades en menys temps.

- Les dades tenen varietat de formats i això requereix nous models de dades per facilitar la inserció, la consulta i el processament de dades de qualsevol tipus i estructura. D'aquesta manera, apareixen els nous models de bases de dades *NoSQL*, que utilitzen estructures de dades diferents a les del model relacional i que permeten tractar més eficientment tipus de dades heterogènies o molt relacionades.

- Es requereix rapidesa per processar les dades. Tot i la seva quantitat, en ocasions es requereixen respostes ràpides. Per exemple, un cercador web no seria gens pràctic si tornés els resultats d'una consulta una hora (o inclús un minut) després d'haver-la realitzat.

Això va empènyer a aquestes empreses a construir les seves pròpies tecnologies per poder continuar amb el model de negoci que ells mateixos havien creat i, al mateix temps, van promoure el desenvolupament de sistemes de codi obert per, d'aquesta manera, poder-los utilitzar sense haver de pagar el cost associat a les llicències que feia gairebé inviable el sistema distribuït.

## 2.5. Del BI tradicional al Big Data Analytics

Disposar d'una gran quantitat de dades de per sí no aporta valor. El valor de les dades rau en analitzar-les i interpretar-les de forma correcta, no el la seva generació o acumulació. *McKinsey Global Institute* en el seu informe "*Big data: The next frontier for innovation, competition, and productivity"* [9], argumenta que les dades s'estan convertint en un factor de producció, tal com ho és el capital físic o humà.

En l'analítica de negoci es té com objectiu principal fer prediccions i/o descobrir tendències, sobre certes característiques d'una població, per ajudar a prendre decisions que repercuteixin de manera positiva en el negoci.

Big Data i Business Intelligence (BI) tradicional són dues tecnologies que permeten gestionar i crear coneixement mitjançant l'anàlisi de les dades per donar suport a la presa de decisions, però hi ha diferències entre Big Data i BI tradicional , i és que difereixen tant en la manera en què ho fan, com en el tipus de dades que analitzen.

En la primera part d'aquest apartat s'entra en la necessitat i aportacions d'analitzar totes aquestes noves fonts de dades i en la segona part es descriuen les limitacions dels sistemes BI tradicionals per fer front a aquesta necessitat.

### *2.5.1. Aportacions de l'analítica de Big Data*

L'anàlisi de dades massives permet a les organitzacions crear nou valor basat en coneixement empíric d'una determinada població fent viable el modelar amb alta precisió, per exemple, el comportament social o mecànic de diferents agents (persones, màquines, respectivament). Això permet obtenir majors

nivells d'optimització en la reducció de costos, en innovació de productes, serveis i processos.

En el camp de l'Estadística, la mida de la mostra és un factor important a tenir en compte. Com més gran sigui la mostra, més exacta serà l'estimació resultant i la prova hipotètica es realitzarà amb millor criteri. Així, si la mostra abastés tota la població, els resultats obtinguts serien precisos.

Amb un sistema Big Data, és te la capacitat de recollir més dades (mostres) i la capacitat de processar més quantitat de dades en menor temps. Així resulta possible analitzar dades que en principi no semblaven prou rellevants com per ser analitzades o es descartaven per la dificultat d'integrar-les. D'aqueta manera, es poden arribar a utilitzar mostres que s'aproximen molt més al total de la població que amb els sistemes d'analítica tradicionals. D'aquesta manera, apareix un canvi de paradigma on la correlació substitueix la causalitat.

Al igual que els sistemes BI tradicionals, es podria respondre a: "què va passar", "què està passant" i "què passaria si", però des d'un punt de vista estadístic, no causal, en el que no es busca l'explicació del fenomen, sinó solament el descobriment del fenomen en si. Aquest canvi de paradigma, fa que els sistemes analítics de Big Data es centrin en "quins" aspectes afecten a la presa de decisió i no en el "per què" afecten aquests aspectes.

Aquests fets eleven l'anàlisi estadística, de Big Data, a nous nivells d'eficàcia que, en l'àmbit empresarial, provoca que es produeixi una pressió tecnològica com a conseqüència de l'anàlisi massiva de dades i del conseqüent risc de perdre competitivitat si aquestes dades no s'exploten adequadament quan la resta d'empreses competidores del sector sí que ho fan.

Entre les diferents tendències apuntades per experts de consultores i centres de recerca, per al 2015 destaca el Big Data, tecnologia en la qual es preveu que les companyies focalitzaran bona part de les seves inversions. L'enquesta anual sobre Big Data que realitza *Gartner*[10] assenyala que hi ha un creixent interès i inversió en aquesta tecnologia per part de les empreses. L'informe revela que el 73% dels enquestats ha invertit o té plans per invertir en Big Data durant aquest any.

Els governs i administracions públiques també es poden aprofitar d'aquest potencial, poden utilitzar les dades per a proporcionar un millor servei i abordar problemes relacionats amb la sanitat, l'ocupació, la prevenció de catàstrofes i el terrorisme. En els EE.UU s'ha utilitzat per predir terratrèmols, les taxes d'atur, per reduir índexs de criminalitat i per a seguiment d'epidèmies. En concret, en el sector de sanitat, *McKinsey [9]* estima que Big Data podria estalviar entre de 300 i 450 milions de dòlars a l'any en els EE.UU.

### *2.5.2. Limitacions de Business Intelligence tradicional*

La metodologia tradicional de BI es basa en el principi d'agrupar totes les dades empresarials rellevants per el seva anàlisi, en general en un sistema *Data Warehouse*. Normalment aquestes dades són incorporades al sistema en mode *off-line* i s'emmagatzemen de forma estructurada en una RDB.

Per analitzar les noves fonts de dades, el BI tradicional té varis inconvenients que deriven de manera directa de les limitacions de les RDB. En primer lloc, dificultats en estructurar dades no estructurades, i en segon lloc, dificultats per realitzar anàlisis de dades en temps real. Es descriuen amb més detall a continuació:

**1) Dificultats per estructurar dades no estructurades**

Els sistemes de BI tradicionals poden treballar amb un gran nombre d'orígens de dades diferents, però al estar basats en sistemes relacionals, les dades han de ser estructurades. Quan dades a incorporar no són estructurades, calen processos que les preparin, interpretin i les integrin amb la resta de les dades. Aquestes tres passes es duen a terme amb els processos coneguts com ETL (en anglès, "*Extraction, Transformation and Load"*).

Al estructurar orígens de dades no estructurades es pot produir una pèrdua d'informació, ja que només s'extreuen i s'emmagatzemen les qüestions que prèviament han estat considerades rellevants. En el cas d'estructurar orígens de dades semiestructurades també es complica el procés de càrrega, ja que obliga a afegir tantes excepcions com possibilitats de variació tinguin les dades.

**2) Dificultats per realitzar l'anàlisi en temps real**

En el cas d'aplicar de funcions estadístiques avançades o tècniques d'intel·ligència artificial, en general es requereixen implementacions a mida, que impliquen: l'extracció de les dades d'interès, emmagatzematge intermedi d'aquestes dades, aplicació dels càlculs sobre aquestes dades i emmagatzematge del resultat. Aquestes funcions, no s'executen dins la base de dades i al haver de passar per diferents processos de forma seqüencial, es genera una potencial pèrdua de rendiment i temps en el cas de grans volums.

En un sistema *Data Warehouse* abans de poder analitzar les dades han de passar per diferents processos ETL, que permeten transformar, normalitzar i carregar les dades al *Data Warehouse*, mantenint, a més, certs criteris de qualitat de dades. Un altre aspecte que cal tenir en compte és la necessitat d'accelerar l'accés a les dades més freqüents; normalment movent les dades a unitats de memòria més ràpides i generant noves estructures. Aquests processos són costosos en temps i en recursos.

## 2.6. BI o Big Data Analytics

Business Intelligence (BI) és un sistema que analitza les dades de l'empresa, normalment estructurades en un *Data Warehouse*, i mostra com està funcionant el negoci en les seves diferents àrees per poder prendre les millors decisions. El Big Data, per la seva banda, recull dades de diferents fonts (tant internes com externes), de volum il·limitat, amb independència de la seva estructura, i proporciona una anàlisi que permetrà avançar-se a les tendències del mercat .

En moltes ocasions, les solucions tradicionals de BI no són suficients per extreure la informació d'aquestes dades, el que pot fer pensar que Big Data substitueix al BI tradicional, ja que ofereix una anàlisi més profund i una visió més global. No obstant això, la implementació de *MapReduce* segueix sent costosa i requereix de moltes àrees de coneixement.

Per la seva banda el BI aporta a l'usuari una forma d'explotar les dades més coneguda i estructurada. Els elements del BI com els *Dashboards*, els informes o les mètriques de rendiment poden ser molt importants a l'hora d'oferir anàlisis fiables, principalment en organitzacions que porten temps utilitzant-les.

En aquest sentit Big Data es pot implementar com una nova font de dades del BI, com una manera d'enfocar una solució que fins al moment era inabastable per les eines BI tradicionals. D'aquesta manera, l'analítica avançada de Big Data i el BI són perfectament complementaris[11].

## 2.7. Definició de Big Data

Es difícil trobar una definició rigorosa del que és Big Data, per una part perquè és un concepte relativament nou i que es troba en evolució. Per altra part, la definició més acceptada no s'implementa a partir del que és, sinó a partir de les característiques de les dades que pretén analitzar. Segons Kenneth Cukier, autor del llibre '*Big Data. La revolución de los Datos Masivos'* [12], es tracta de fer coses a partir de l'anàlisi d'immenses quantitats d'informació, que simplement no són possibles amb volums més petits.

De forma més general, es denomina Big Data a la gestió i anàlisi d'enormes volums de dades que no poden ser tractats de manera convencional, ja que superen els límits i capacitats de les eines de software habitualment utilitzades per la captura, gestió i processament de les dades.

Des del punt de vista més tecnològic, Big Data és un sistema genèric que engloba infraestructures, tecnologies i serveis, i que integra els diferents components depenent de la quantitat, tipus i relació de les dades, models i algoritmes a executar, amb la finalitat de donar suport i aconseguir un millor rendiment al processament d'enormes conjunts de dades.

Des de el punt de vista del negoci, el *Big Bata* és la possibilitat de comptar amb un nombre enorme de dades que fan reduir al màxim els costos de transacció

en la presa de decisions de les organitzacions, a través de tenir un nombre gran d'observacions empíriques.

L'objectiu de Big Data, al igual que els sistemes analítics convencionals es convertir les dades en informació per l'ajuda a la presa de decisions, fis i tot en temps real. Però tal i com succeeix amb qualsevol altre model de negoci, el factor clau per obtenir beneficis de Big Data, no depèn de la capacitat tecnològica sinó de la capacitat humana per realitzar la correcta interpretació de la informació que permetrà obtenir el valor de l'anàlisi, és a dir, el coneixement.

### *2.7.1. Les Tres Vs*

Les diferencies entre les aplicacions analítiques i de gestió tradicional i els nous conceptes Big Data, s'associen en la majoria d'articles de referència a les famoses tres '*Vs*' del Big Data (*Volum, Varietat i Velocitat*) i que han estat àmpliament adoptades i adaptades. Aquestes tres característiques foren definides per Doug Laney (2001) de l'actual *Gartner, Inc.* en el seu article '*3D Data Management: Controlling Data Volume, Velocity, and Variety*'[13] quan va identificar la tendència cap a l'anàlisi massiu de dades.

**1) Volum**

El concepte de volum es dona quan la mida de les dades supera la capacitat del software habitual per tractar-lo o gestionar-lo. Aquest concepte no es estàtic ja que els nous dispositius d'emmagatzemament cada cop permeten tractaments de volums més grans. Tot i així, en parlar de grans volums es fa referència a l'ordre de Terabytes, Petabytes o superiors. En termes de bytes:

*Gigabyte* = $10^9$ = *1.000.000.000*
*Terabyte* = $10^{12}$ = *1.000.000.000.000*
*Petabyte* = $10^{15}$ = *1.000.000.000.000.000*
*Exabyte* = $10^{18}$ = *1.000.000.000.000.000.000*

**2) Varietat**

El concepte de varietat es refereix a la inclusió d'altres tipus de fonts de dades diferents a les tractades de forma tradicional. Informació provinent de xarxes socials, dispositius electrònics connectats, sensors, etc. Aquesta informació es pot trobar en diferents formats: estructurada, tal i com es troba en un una RDB; semiestructurada o sense cap estructura. En parlar de Big Data es refereix més concretament als no estructurats. La gestió d'informació no estructurada precisa d'una tecnologia diferent que permet prendre decisions basades en informació que te importants graus d'inexactitud.

**3) Velocitat**

El concepte de velocitat es refereix a la rapidesa amb que les dades es creen, processen i analitzen per generar resultats. La velocitat afecta la latència: el temps d'espera entre el moment en què es creen les dades, el moment en què es capten i el moment en què estan accessibles. Avui en dia, les dades es generen de forma contínua a una velocitat a la qual els sistemes tradicionals els resulta impossible captar-les, emmagatzemar-les i analitzar-les. Per als processos en els quals el temps és fonamental, com ara la detecció de frau en temps real o el màrqueting "instantani" multicanal, certs tipus de dades s'han d'analitzar en temps real perquè resultin útils per al negoci.

### *2.7.2. Les altres Vs*

Tot i així, en base a l'experiència adquirida per empreses pioneres amb Big Data, s'ha ampliat la definició original, afegint noves característiques com són la Veracitat i la Viscositat o Valor.

**4) Veracitat**

La *Veracitat*[14] es refereix tant a la qualitat de la dada com a la seva predictibilitat i rellevància. La varietat afecta la veracitat, pel que aquesta darrera és la variable menys uniforme al llarg dels diferents tipus de dada, ja que porta implícit el biaix, el soroll i l'alteració dels mateixos. Esforçar-se per aconseguir unes dades d'alta qualitat és un requisit important i un repte fonamental de Big Data, però fins i tot els millors mètodes de neteja de dades no poden eliminar la incertesa inherent d'algunes dades, com el temps, l'economia o les futures decisions de compra d'un client. La necessitat de reconèixer i planificar la incertesa és una dimensió de Big Data que sorgeix a mesura que els directius intenten comprendre millor el món que els envolta.

**5) Viscositat o Valor**

Finalment, s'afegeix la viscositat o valor que fan referència, en un sentit similar, a la major o menor facilitat per correlacionar les dades i a la importància d'aquestes per el negoci o organització. La capacitat de recuperar informació útil de fonts heterogènies i correlacionar les dades entre elles pot ser una tasca complexa, generalment mitjançant sofisticats algoritmes, per aconseguir generar coneixement que pot resultar molt rellevant per la presa de decisions. Tant que ja s'ha definit un nou perfil professional[15], el científic de dades; cada cop més demandats per les empreses.

## 2.8. Científic de dades

El científic de dades[16] correspon a un perfil professional multidisciplinari, que ha de comptar amb un nivell d'estudis alt i que ha de tenir unes determinades característiques i habilitats; es tracta d'un conjunt de competències i coneixements que pot ser difícil de trobar en un sol individu. Per aquest motiu, en parlar de científic de dades es pot pensar en un equip d'experts que treballi de manera conjunta per donar un valor afegit al producte final d'una empresa. Aquest equip de treball ha de comptar amb experts en enginyeria de dades, mètode científic, matemàtiques, estadística, computació avançada, visualització i experts en els diferents àmbits d'especialitat.

Per agrupar tot aquest coneixement que s'ha concentrant al voltant del terme de Big Data ha emergit el concepte '*Data Science'* o 'Ciència de les Dades' que engloba certes àrees de la Intel·ligència Artificial, Estadística i Mineria de dades.

# 3. Ciència de les Dades

La Ciència de les Dades (en anglès, *Data Science*) és el terme nou que apareix amb la popularització de Big Data. El seu punt de partida és el mateix que Big Data: ingents quantitats de dades que no poden tractar-se amb les tècniques convencionals de tractament de dades.

Emmarca coneixements, sobre fonts de dades i la seva organització, tècniques de processament massiu de dades, tècniques de tractament de dades, repositoris especialitzats d'emmagatzematge de Big Data, tècniques i metodologies d'anàlisi de dades i de visualització d'aquestes anàlisis. Així mateix inclou una visió horitzontal i complementària en coneixements sobre el maquinari, programari, sistemes operatius, *middleware*, *frameworks* i aplicacions en general per poder explotar les dades massives. En aquest capítol s'introduiran moltes d'aquestes tècniques com a base per anar introduint les restants en el transcurs del treball.

Les àrees de *Data Mining*, *Machine Learning* i *Processament de Llenguatge Natural* també queden albergats sota l'epígraf de *Data Science*.

## 3.1. Fonts de dades. Taxonomies

Seleccionar i mantenir les fonts d'informació, es una tasca essencial en els projectes Big Data. Totes les fonts de dades necessàries han de ser estudiades i catalogades per tal que puguin respondre als objectius i necessitats del sistema.

La Taxonomia és la ciència de la classificació i així mateix s'utilitza com un sinònim de classificació. Originalment s'utilitza en el món de la biologia, però el seu ús s'ha estès a la resta d'àmbits del coneixement. Davant l'explosió de fonts d'informació a Internet s'ha fet especialment rellevant la classificació dels seus camps, amb l'objectiu d'harmonitzar tot el possible els continguts i poder així creuar i unir diferents fonts d'informació gràcies a l'estandardització que proporciona l'ús de taxonomies.

Existeixen nombroses taxonomies utilitzades en l'actualitat, que són promogudes i mantingudes per diferents organismes nacionals i internacionals i en molts àmbits diferents. Per exemple, en l'àmbit sanitari, la CIM-9 i més recentment CIM-10 és l'acrònim de la Classificació Internacional de Malalties[17], versió 9 i 10, respectivament. Aquest acrònim es correspon a la versió en català de la (en anglès) ICD, '*International Statistical Classification of Diseases and Related Health Problems'*. Determina la classificació i codificació de les malalties i una àmplia varietat de signes.

En els casos de trobar-se davant fots classificades amb diferents taxonomies, caldrà comprovar si existeixen passarel·les que estableixin correspondència entre elles. En cas de no existir, caldrà dissenyar un mecanisme de traducció entre les taxonomies.

La integració tècnica de les fonts pot ser més o menys complicada, depenent de les diferents circumstàncies i de com s'estructuren les fonts de dades. En general, les eines de ETL dels paquets de programari de Business Intelligence són suficients per realitzar la integració fiable d'una font d'informació estructurada.

Per al tractament de fonts no estructurades s'estan utilitzant diferents tècniques, disciplines i metodologies com:

- *Scraping*
- *Linked Data*
- Processament de Llenguatge Natural
- *Machine Learning*
- Intel·ligència Artificial, Estadística, Investigació Operativa i *Data Science* en general

## 3.2. Tècniques per la integració de dades

En aquest apartat es fa una introducció a diferents tècniques per la integració de dades en un sistema Big Data*,* La majoria d'aquestes ja utilitzades en Busines Intelligence i altres adaptades. Aquestes són l'*Scraping*, els processos ETL/ELT i la federació de dades. A continuació es descriuen les tres.

### 1) Scraping

*Scraping* és un conjunt de tècniques que tenen com objectiu extreure informació de dades no estructurades de diferents formats com HTML, XML o col·leccions de documents i transforma-les en dades estructurades que poden ser emmagatzemades i analitzades en una base de dades relacional. Normalment aquests programes s'utilitzen per simular la navegació d'una persona per Internet ja sigui utilitzant el protocol HTTP, o incrustant un navegador a una aplicació amb l'objectiu d'obtenir informació útil que no està disponible en cap altre forma.

Per a això es requereixen aplicacions específiques, anomenades *Webbots* o *Bots*, que automatitzen la interacció amb el lloc web en la informació estem interessats. Els *Bots* realitzen diverses funcions, destacant la funció de navegació per la pàgina web, seguiment dels enllaços i la de lectura dels continguts. En general, requereixen desenvolupaments a mida i hi ha diferents enfocaments que hauran de ser seleccionats en funció del problema concret a resoldre, la capacitat tècnica de l'equip encarregat d'això i la infraestructura disponible.

### 2) ETL/ELT

La primera etapa dels projectes de Business Intelligence és tradicionalment l'ETL, és a dir, l'Extracció de Dades de fonts, la transformació d'aquestes dades i la càrrega de les dades transformades en una base de dades relacional.

La tendència Big Data es convertir els processos ETL en processos ELT[18], és a dir, després de l'extracció de les dades de la font no es fa cap tipus de transformació sinó que es carreguen totes les dades en el sistema, per més endavant, poder fer totes les anàlisis i transformacions que siguin pertinents. En definitiva, fins a l'arribada de Big Data les dades es quedaven a les fonts d'informació, integrant només la informació necessària en els sistemes. Amb Big Data, la cadena '*Dada > Informació > Coneixement'* es troba de forma completa en el sistema. La raó fonamental és habilitar que a posterior sigui viable realitzar tant prediccions basades en dades històriques com respondre a casos d'ús que han de ser definits en el futur.

Tot i així, en alguns casos, el processos ETL poden ser clau per incorporar dades externes a les bases de dades. Aquets són els processos de *Neteja de Dades* (en anglès '*data cleansing*'), i els d'*Enriquiment de Dades* (en anglès '*data enrichment*').

Amb el procés de *Neteja de Dades* s'intenta proporcionar major qualitat a les dades. Si el sistema gestiona dades errònies, incompletes, duplicades o poc veraces, els resultats que es generaran contindran contradiccions, anomalies, problemes de consistència i irregularitats que portaran als usuaris a prendre decisions incorrectes ja que estaran basades en dades amb una qualitat inferior a la necessària.

L*'Enriquiment de les Dades* s'ha convertit en un procés clau en els projectes Big Data. Freqüentment les dades disponibles o amb els quals s'ha treballat habitualment en l'organització no són suficients per donar solució als casos d'ús i objectius de negoci plantejats. Una solució consisteix a acudir a la gran quantitat de noves dades disponibles a Internet gràcies al moviment *Open Data* o directament a empreses que s'han especialitzat en obtenir, mantenir i posar dades a disposició de tercers.

**3) Federació de les dades**

Una altra tendència per la integració és la *Federació de Dades*, (en anglès '*Data Federation*'). Funciona fent us d'una base de dades federada que intermèdia amb altres bases de dades de forma transparent a l'usuari, oferint-li una única interfície d'accés a totes les dades. No caldrà per tant traslladar a un únic repositori totes les dades, sinó que cada base de dades roman autònoma i es programa a la base de dades federada com accedir a les dades de cadascuna de les altres bases de dades.

## 3.3. Emmagatzemament i gestió de dades.

Per allotjar la gran volum i varietat de dades es requereixen nous tipus de bases de dades, el anomenades *NoSQL* són els nous contenidors d'informació especialment preparats per als tipus de processament necessaris. Aquestes, a més de dades i informació, gestionen el coneixement en Ontologies, que són reflex de la 4ª V, la Veracitat. Els sistemes de fitxers distribuïts i paral·lels són la base dels sistemes Big Data.

### *3.3.1. Sistemes de fitxers distribuïts i paral·lels*

Els sistemes de fitxers distribuïts permeten incorporar milers d'ordinadors independents que es converteixen en nodes independents d'una mateixa xarxa. Gestionen diferents dispositius en diferents nodes de forma transparent a usuaris i aplicacions oferint serveis amb les mateixes prestacions que si fos un sistema de fitxers centralitzat. En cas que un node falli, el sistema de fitxers gestiona automàticament la situació. Per tant, els sistemes distribuïts destaquen per la seva escalabilitat.

Els sistemes de fitxers paral·lels són un grup específic dins dels sistemes distribuïts que es caracteritzen la distribució de dades entre múltiples dispositius d'emmagatzematge i l'accés paral·lel als mateixos. Es popularitzen davant les següents necessitats:

- Transportar una gran quantitat d'arxius sobre la xarxa pot causar baixes prestacions causa de l'alta latència, colls d'ampolla a la xarxa, alta escalabilitat i sobrecàrregues.
- La necessitat creixent de les aplicacions de gestionar repositoris massius de dades.
- La manca de creixement de l'ample de banda i la latència als discs enfront del creixement enorme de la seva capacitat.

Tant per els sistemes distribuïts com per els paral·lels es compta amb que les fallades en el maquinari i en el programari són norma, i no l'excepció. Per això aquests sistemes compten amb mecanismes de replicació que gestionen la situació per no perdre les dades.

### *3.3.2. Bases de dades Bigdata*

A les bases de dades Big Data se'ls agrupa sota el concepte de bases de dades *NoSQL*. Com altres paradigmes, el concepte no és nou, té les arrels en les bases de dades en xarxa i jeràrquiques de l'últim quart del segle XX.

L'emmagatzemament distribuït no relacional, és un dels fonaments del Big Data. Els sistemes d'emmagatzematge Big Data manegen de forma distribuïda una quantitat de dades de l'ordre de petabytes, en clústers de milers de servidors *commodity* barats, amb un rendiment molt alt, complint el principi CAE, o sigui eficient en costos, amb alta disponibilitat i ampliable elàsticament (en anglès *'Cost-Efficiency', 'High Availability', 'Elasticity'*).

Un aspecte que cal destacar com a possibles inconvenients és que a diferència de les Bases de Dades Relacionals (RDB), les *NoSQL* no tenen un llenguatge de consulta en comú (com SQL) i tampoc compleixen totes les propietats ACID (atomicitat, consistència/integritat, aïllament i durabilitat/persistència). Les bases de dades *NoSQL* compleixen les propietats BASE que són les sigles de "*Basic Availability*", "*Soft-state*" i "*Eventual consistency*". Aquestes propietats signifiquen que després de cada transacció, les bases de dades *NoSQL* estaran en un estat consistent, que es poden lliurar dades obsoletes i que es

permeten donar respostes aproximades. En aquest sentint cal avaluar si el model a implementar pot permetre aquestes situacions.

Algunes de les raons que poden aconsellar l'elecció d'una base de dades *NoSQL* en lloc d'una RDB són:

- Facilitat a l'hora de gestionar varietat de dades semiestructurades, per exemple quan en cada inserció de dades la informació a emmagatzemar té camps diferents.
- Necessitat de gestionar i mantenir grans volums de dades (terabytes a petabytes) especialment en pics d'ús del sistema.
- Problemes d'amplada de banda per ingerir els fluxos de dades entrants: quan tenim pics d'ús del sistema que provoquen problemes operatius, per exemple quan les dades arriben a una velocitat superior a la que podem gestionar
- Quan es presenten problemes d'escalabilitat amb les RDB tant tècniques com econòmiques (costos de llicències, replicació en diferents centres de dades, *cloud computing*, etc.).
- Quant la complexitat de les consultes és superior a la es pot gestionar amb RDB. Aquesta situació es millora amb un tipus de bases de dades relacionals que funcionen en paral·lel però una BD NoSQL pot ser millor solució.
- Hi ha alta concurrència en les consultes a la base de dades o són molt intensives en l'ús de la CPU.
- Quan s'integren dades molt interconnectades, dades massives estructurades i dades no estructurades amb alguna relació.

A l'apartat '4.3. Bases de Dades *NoSQL*' s'analitzen amb més detall.

### *3.3.3. Ontologies*

Les ontologies són descripcions de coneixement, esquemes conceptuals en dominis d'informació concrets. Aquests esquemes permeten classificar el coneixement i raonar sobre ell de forma automatitzada.

Una ontologia és una especificació formal i explícita d'una conceptualització compartida, on la semàntica de la informació es representa mitjançant objectes, relacions i propietats que els caracteritzen, en un llenguatge que sigui comprensible per als ordinadors, és a dir, un llenguatge formal. Per tant, una ontologia és un model de coneixements consensuat en un determinat domini i que és reutilitzable en diferents aplicacions.

## 3.4. Tècniques de processament i anàlisi

El Processament i anàlisi de Big Data es portat a terme per un sistema de processament distribuït que aprofita tecnologies tradicionals com la programació funcional, el *Machine Learning*, el Processament de Llenguatge Natural, i un grup d'àrees de coneixement que agrupem sota el concepte de la

*Data Science* i la Intel·ligència Artificial. Aquests grups de processos fan referència a la 5ª "V", de Valor o Viscositat.

### *3.4.1. MapReduce*

Fins fa pocs anys no s'ha disposat d'una infraestructura de maquinari i programari que fes viable des d'un punt de vista tècnic i econòmic l'aplicar aquest tipus de tècniques a quantitats massives de dades. L'elevat temps de computació necessari feia inviable l'aplicació del paradigma funcional al tractament massiu de la informació en un temps acceptable. Els nous sistemes distribuïts si permeten aquest tipus de solucions mitjançant la paral·lelització en clústers d'ordinadors estàndard de baix cost.

MapReduce és un model de programació utilitzat per desenvolupar aplicacions paral·leles a gran escala per el processament de grans quantitats de dades. Va ser implementat per primera vegada per Google en 2004 i des de llavors s'ha convertit en una eina important per a la computació distribuïda. És especialment adequat per a tractar grans conjunts de dades en clústers de computadors, ja que està dissenyat per tolerar fallades de maquinari.

Mentre que MapReduce no és un concepte revolucionari, s'inspira en solucions dissenyades fa decennis similars a les que es poden trobar en LISP i altres llenguatges de programació funcionals, ha contribuït a estandarditzar les aplicacions paral·leles i, tot i que la seva interfície és senzilla, ha demostrat ser prou potent com per a resoldre una àmplia gamma de problemes del món real que poden ser solucionats aplicant aquest model i específicament molts relacionats amb Big Data.

De forma general, MapReduce divideix el treball de cada aplicació en petits càlculs en dos passos principals[19], *Map* i *Reduce*. L'entrada està formada per un conjunt de parells clau/valor, que es processen utilitzant la funció *map()* definida per l'usuari per generar un segon conjunt de parells clau/valor intermedis. Els resultats intermedis són després processats per la funció *reduce()*, que combina els valors per la clau.

#### ***Arquitectura escalable i fiable***

MapReduce ofereix una alta escalabilitat gràcies a la divisió del treball en fragments més petis, on cada un d'aquests és executat en un node del sistema distribuït. Els treballs s'envien a un node principal (*Master*), que és l'encarregat de gestionar l'execució d'aplicacions en el clúster. Després de la presentació d'un treball, el node principal inicialitza el nombre desitjat de tasques o unitats de treball més petites i els posa a funcionar en nodes de treball (*Worker*). En primer lloc, durant la fase *Map*, els nodes llegeixen i apliquen la funció *map()* a un subconjunt de les dades d'entrada. La sortida parcial de *map()* s'emmagatzema localment a cada node i es proporciona als nodes de treball que executen la funció *reduce()*. Tant els arxius d'entrada (*Imput data*) com els de sortida (*Results*) es solen emmagatzemar en un sistema d'arxius distribuït.

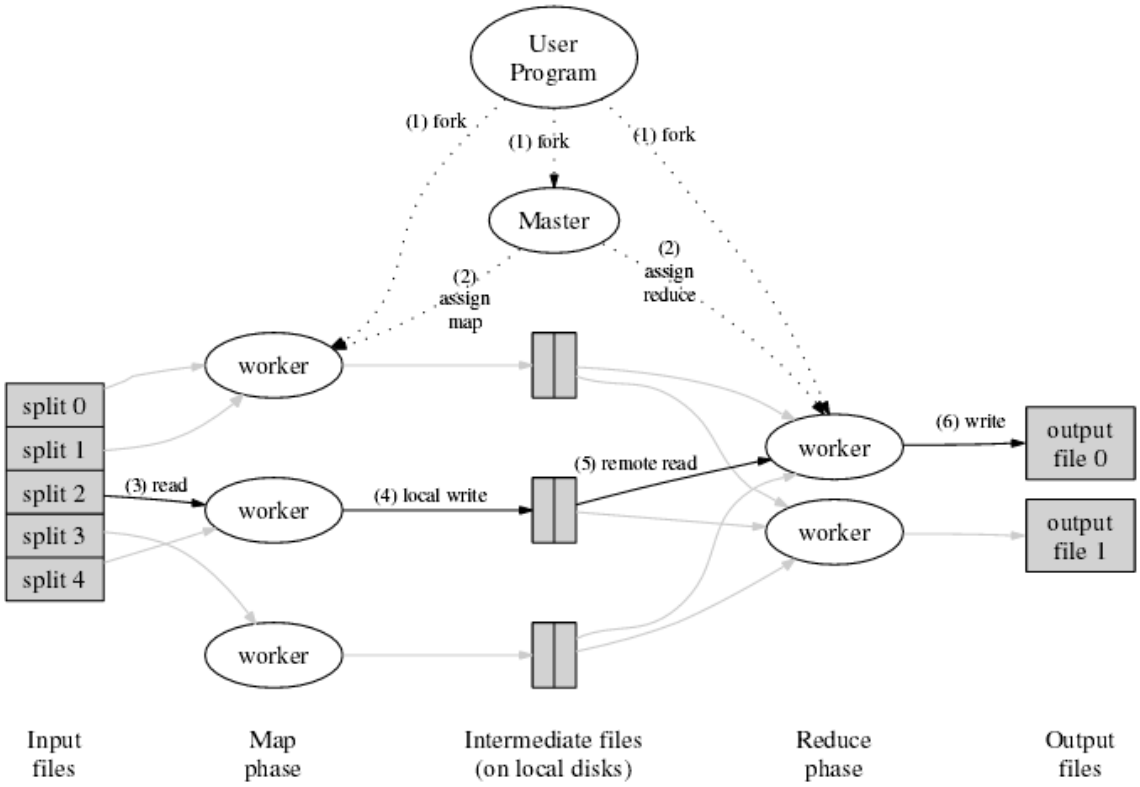


**Figura 3. Arquitectura MapReduce**

La implementació de MapReduce també ofereix un alt nivell de fiabilitat, garantint que les operacions *Map* i *Reduce* són deterministes respecte a les seves entrades (és a dir, que sempre produeixen els mateixos resultats per a un conjunt donat d'entrades), així, el treball en conjunt de MapReduce produirà el mateix resultat que una execució seqüencial de l'aplicació, fins i tot davant fallades de nodes.

Per tal de proporcionar la fiabilitat, el node principal comprova periòdicament que un treballador està actiu i duu a terme la seva operació prevista. D'aquesta manera, si un node treballador no pot lliurar o completar la unitat de treball que se li ha assignat, el node principal és capaç d'enviar aquest treball a un altre node.

### *Casos d'us i exemple*

MapReduce es pot utilitzar en molts tipus diferents d'aplicacions, des d'eines d'ajuda molt simples que són part d'un entorn més ampli, a programes més complexos que poden implicar múltiples execucions de MapReduce encadenats.

En la taula següent es mostra un conjunt d'exemples d'aplicacions comuns i la forma en què es durien a terme per les funcions *Map* i *Reduce*. Els passos compartits (Pas inicial i Pas intermedi) en el còmput realitzat pel marc de treball MapReduce es mostren també per aportar més completesa. Més detalls d'aquests exemples es poden trobar en *Dean i Ghemawat* (2004)[20].

**Taula 1. Aplicacions i fases MapReduce**

| Aplicació | Pas inicial → | Fase Map → | Pas intermedi → | Fase Reduce |
|---|---|---|---|---|
| Recompte de paraules | Partició de les dades en blocs de mida fitxa | Per cada ocurrència de paraula en la partició genera el parell <paraula,1> | Combina/ordena tots els parells clau/valor d'acord amb la seva clau intermèdia | Per cada paraula del conjunt intermedi, conta el nombre de 1s |
| Cerca de patrons | | Treu el resultat si aquesta coincideix amb el patró donat | | Buit |
| Ordenació distribuïda (depèn en gran mesura del pas intermedi) | | Per cada entrada, treu el parell clau/valor per ser ordenat | | Buit |
| Indexació invertida | | Analitza els documents i treu el parell <paraula, ID document> on sigui que hi hagi aquesta paraula | | Per cada paraula, produeix una llista d'IDs de document ordenada |

L'aplicació de recompte de paraules és la que exemplifica MapReduce en el document original i des de llavors s'ha convertit en l'exemple principal[21] per presentar com funciona aquest model de programació. L'objectiu d'aquesta aplicació és aconseguir la freqüència de les paraules en una col·lecció molt gran de documents.

Per portar a terme la tasca de recompte de paraules amb el model MapReduce cal implementar les dues funcions que a grans trets i en pseudocodi tindria la forma següent:

```
map(String clau, String valor):
        // clau: (no s'utilitza per aquest exemple)
        // valor: cadena d'entrada
        for each paraula w in valor:
                Emet_Intermedi(w, 1);
```

La funció *map()* és simple: s'agafa la cadena de l'entrada (que correspon a una partició de les dades), i per cada paraula emet un parell clau/valor <paraula, recompte>, on el recompte és el recompte parcial i és sempre 1.

```
reduce(String clau, iterator valors):
        // clau: la paraula
        // valors: la llista de recomptes de la paraula
        int resultat = 0;
        for each v in valors:
                resultat += v;
        Emet(clau, resultat);
```

La funció *reduce()* agafa els parells <clau, llista(valors)> i recorre tots els valors per obtenir la suma del conjunt de tots els recomptes emesos per una paraula concreta.

Cal destacar que en alguns casos hi pot haver una repetició significant en les claus del pas intermedi produïdes per cada tasca *map()*, i com que la funció *reduce()* és commutativa i associativa, es podria especificar una funció de combinació parcial que seria executada en cada node que realitza la tasca *map()*. D'aquesta manera, es podrien reduir la quantitat de parells que s'enviarien a la funció *reduce()* i en conseqüència, augmentar la velocitat de processament en certes operacions MapReduce.

A tall d'exemple ens podem imaginar el següent. El node principal rep un treball de recompte de paraules d'un document *x* amb els següent contingut "Hola Món Hola BigData". El node principal el particiona i el posa a treballar en dos nodes de forma simultània. Un executarà *map*(x, "Hola Món") i l'altre *map*(x, "Hola BigData"). Cada un d'aquests emetrà un resultat parcial diferent, el primer (<Hola, 1>, <Món, 1>) i el segon (<Hola, 1>, <BigData, 1>).
En el pas intermedi, el resultats parcials són combinats agrupant per la clau, formant els parells <Hola, (1, 1)>, <Món, (1)>, <BigData, (1)>. Aquests parells després són processats per la funció *reduce()* que produeix el resultat final <Hola, 2>, <Món, 1>, <BigData, 1>.

MapReduce és capaç de realitzar funcions molt més complexes que recomptes, cerques, ordenacions i indexació invertida. També és possible encadenar execucions múltiples, utilitzant la sortida d'una aplicació com l'entrada de la següent, permetent cerques i ordenacions distribuïdes, anàlisis de registres, motors de cerca, i fins i tot resoldre alguns problemes de grafs.

### *3.4.2. Machine Learning*

El *Machine Learning*, també conegut com '*Aprenentatge Automàtic*', és una disciplina científica habitualment enquadrada dins l'àrea de la Intel·ligència Artificial, encara que també és possible veure-la enquadrada en l'àrea de l'Estadística.

Sota el concepte de *Machine Learning* s'enquadra un procés pel qual donat un conjunt de dades d'exemple disponibles a partir d'un cas d'ús, objectiu o tasca a realitzar, es dissenya un algoritme que, generalitzant a partir de característiques de les dades de exemple és capaç de resoldre aquest cas d'ús, objectiu o tasca tant per a les dades disponibles com per a altres dades disponibles a posteriori a les quals se li apliqui l'algoritme dissenyat, dins d'un marge d'error considerat com acceptable.

De manera col·loquial, se sol dir que *Machine Learning* és una tècnica mitjançant la qual l'ordinador aprèn dels exemples i l'aprenentatge s'aplica per resoldre els nous casos que sorgeixin.

S'utilitza en aplicacions de molt diferents àmbits: detecció de malalties a partir de dades clíniques, visió artificial, classificació de documentació en general (per

exemple detecció d'*spam*), detecció de frau en el sector bancari, segmentació de clients (per exemple per determinar productes del seu interès o gust), prediccions en borsa, inversió de capitals, detecció d'anomalies en general en les dades (per exemple fallades de xarxa), predicció de preus, traducció automàtica o establiment de rànquings (per exemple en respostes a recerques d'informació), predicció de comportament humà, detecció d'atacs informàtics, identificació de sons de races animals en enregistraments, interfícies home-màquina en neurociències o en jocs, etc.

### *3.4.3. Processament de Llenguatge Natural*

Processament del llenguatge natural, sovint abreujat com PLN (en anglès, NLP, *Natural Language Processing*), es refereix a la capacitat d'un equip informàtic per entendre el llenguatge humà tal com es parla. El PLN és un component clau de la intel·ligència artificial (IA).

L'objectiu del PLN és la d'ajudar als ordinadors a entendre el llenguatge segons el parlen les persones, si fos una realitat, podrien desaparèixer els llenguatges de programació com Java, Ruby o C. Amb el processament del llenguatge natural, els ordinadors serien capaços d'entendre directament a les persones mitjançant llenguatge humà.

Dos són els enfocaments més habituals en l'anàlisi de textos. D'una banda tenim *Machine Learning*, que està basat en mètodes probabilístics, actualment és la tècnica més comú per aquest tipus d'enfocaments Big Data basats en classificacions. D'altra banda tenim els enfocaments lingüístics, basats en la disciplina de la Lingüística Computacional, que són menys utilitzats.

La Lingüística comprèn l'estructura de les frases. L'Anàlisi Lingüístic utilitza el coneixement sobre el llenguatge (gramàtiques, ontologies i diccionaris) el que permet tractar amb l'estructura del llenguatge a tots els nivells: morfològic, sintàctic i semàntic.

Les anàlisis de sentiment o d'opinió en xarxes socials, enquestes, etc. són bons exemples d'aplicacions, on l'anàlisi semàntica de PLN està proporcionant les millors solucions. Però encara s'està molt lluny de fer una comprensió profunda d'un text complex i de poder disposar d'un programari intel·ligent capaç de realitzar pensaments complexos. Tot i així, el seu futur pot ser prometedor.

Ja existeixen diferents funcions de PLN incorporades en programaris que realitzen diferents tasques[22]:

- Separació de frases, etiquetatge gramatical, i anàlisi: el processament del llenguatge natural pot ser utilitzat per analitzar parts d'una oració per comprendre millor la construcció gramatical de la frase, comptar paraules, etc.

- Anàlisi profunda: consisteix en l'aplicació de tècniques avançades de processament de les dades amb la finalitat d'extreure informació específica dels conjunts de grans dades o de múltiples fonts. És

particularment útil quan es tracta de consultes precises o molt complexes amb dades no estructurades i semiestructurades. És utilitzat sovint en el sector financer, la comunitat científica, el sector farmacèutic i les indústries biomèdiques. Cada vegada més, però, l'anàlisi profunda també està sent utilitzat per les organitzacions i empreses interessades en mineria de dades, o en trobar un valor de negoci a partir de conjunts de dades dels consumidors.

- Traducció automàtica: el PLN és cada vegada més utilitzat en programes de traducció automàtica, en la qual una llengua es tradueix automàticament en un altra.

- Extracció d'entitats nominals (*Named entity extraction*): En la mineria de dades, una definició d'entitat amb nom és una frase o paraula que identifica clarament un element, d'un conjunt d'altres elements que tenen atributs similars. Els exemples inclouen noms i cognoms, edat, ubicació geogràfica, adreces, números de telèfon, adreces de correu electrònic, noms d'empreses, etc. l'Extracció d'entitats nominals, de vegades també anomenat reconeixement d'entitats, facilita la mineria de textos.

- Resolució Co-referenciada: En un tros de text, la resolució de la coreferència es pot utilitzar per determinar quines paraules s'utilitzen per referir-se als mateixos objectes.

- Resum automàtic: el processament del llenguatge natural pot ser utilitzat per produir un resum llegible a partir d'un text més gran. Per exemple, produir un breu resum d'un article acadèmic dens.

Els beneficis del processament del llenguatge natural poden ser molt diversos: pot ser aprofitat per a millorar l'eficiència dels processos de documentació, millorar l'exactitud de la documentació, identificar la informació més pertinent de grans bases de dades. Per exemple, un hospital podria utilitzar el processament del llenguatge natural per obtenir dades d'un diagnòstic específic a partir de les notes no estructurades d'un metge i assignar un codi de facturació.

# 4. Tecnologies Big Data

El Big Data ha passat en poc temps de ser una tecnologia innovadora a convertir-se en un mercat i, ara, a transformar-se en una indústria. El sistema més utilitzat en aquesta indústria per oferir capacitats analítiques avançades és *Hadoop*, un programari de codi obert on seu desenvolupament és coordinat per *l'Apache Foundation*. Aquest programari facilita l'emmagatzematge d'informació i respondre a consultes complexes sobre les bases de dades existents amb rapidesa.

Hi ha plataformes que competeixen amb *Hadoop* en l'escenari de Big Data, tot i aquesta segueix sent la més adoptada. Els projectes *Spark* i *Storm*, també de codi obert, avancen a marxes forçades. Altres solucions que suren en el mercat són *HPCC Systems*[23] que disposa d'una versió comunitària i una de pagament amb serveis addicionals, Disco[24] és un *framework* lleuger, de codi obert per a la computació distribuïda basada en el paradigma MapReduce.

En aquest capítol s'estudiarà la plataforma *Hadoop* i els seus principals components, el sistema de processament MapReduce i el sistema de fitxers *HDFS*. També s'aprofundirà en els nous motors de processament *Apache Spark* i *Apache Storm*. Finalment s'analitzaran els principals tipus de bases de dades *noSQL*.

## 4.1. Hadoop

*Hadoop*[25] és el grup de projectes que ha popularitzat Big Data al món. Proporciona una plataforma de codi obert per analitzar i processar Big Data. Esta inspirat en les implementacions de *Google* MapReduce i *Google File Systen* (GFS). És usat d'una manera o una altra per grans empreses que tenen en el seu nucli tècnic el Big Data. Exemples destacats són Twitter, Facebook, LinkedIN, Reddit o Amazon.

L'origen de *Hadoop* es remunta a 2004, quan l'enginyer de programari Doug Cutting, que aleshores treballava a Google, descriu en un document tècniques per manejar grans volums de dades, descompassant-les en problemes cada vegada més petits per fer-los abordables. Poc després va marxar a Yahoo i allí va seguir investigant fins a completar el desenvolupament de la plataforma en 2008.

*Hadoop* és un *framework* que permet processar grans volums de dades a través de servidors *commodity* (de baix cost). El nucli de *Hadoop* sobre el qual la major part dels altres components es construeixen està format per 4 mòduls.

- *Common*: proporciona utilitats i interfícies per donar suport a la resta de mòduls (configuració, la serialització, RPC, etc.).

- *Hadoop Distributet File System (HDFS):* sistema d'arxius distribuït que s'executa en grans clústers i proporciona l'execució distribuïda d'aplicacions i l'accés paral·lel a les dades amb alt rendiment.
- *Hadoop YARN*: planificador de tasques i gestió dels recursos del clúster.
- *Hadoop* MapReduce: *framework* de programari per al processament paral·lel de grans conjunts de dades .

L'arquitectura de *Hadoop* està formada per dues capes on dos d'aquests mòduls són els components essencials, *HDFS* i MapReduce.

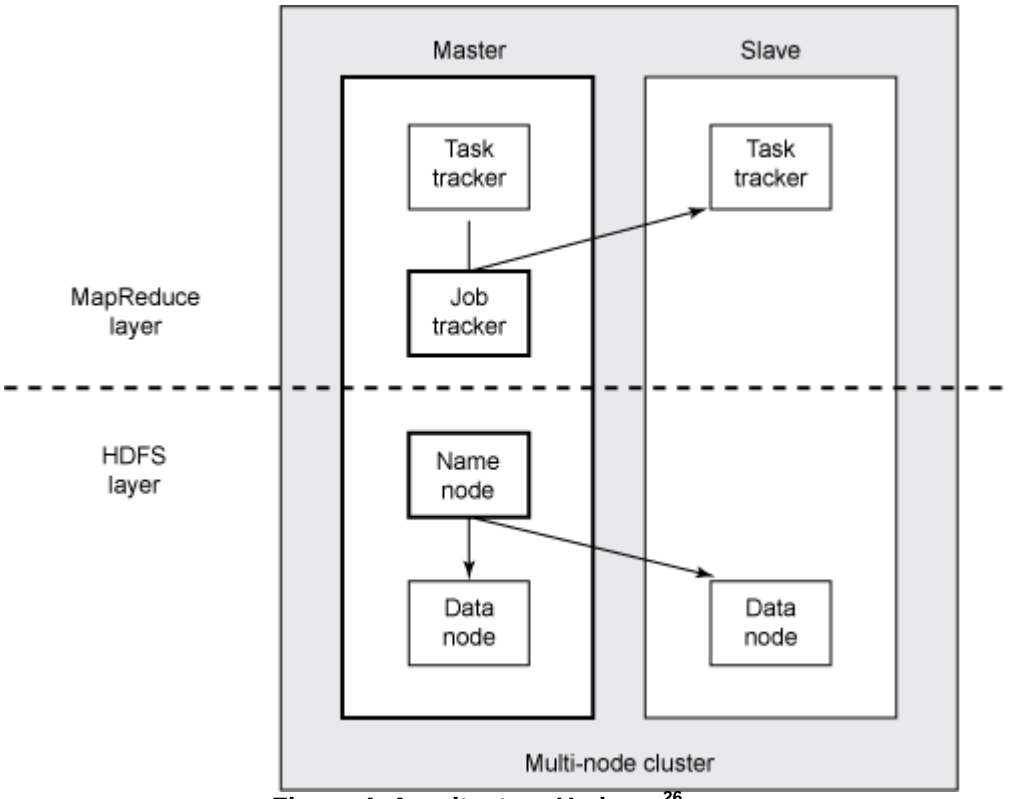


**Figura 4. Arquitectura Hadoop** [26]

La capa *HDFS* permet emmagatzemar grans quantitats de dades, que a més es repliquen i és distribueixen de tal manera que si una màquina deixa de funcionar no es perdin les dades. Si cal afegir més informació s'afegeixen més servidors sense que hi hagi problemes de compatibilitat o reorganització de les dades.

Dos tipus de nodes són els que formen el clúster:

- El *NameNode*: Només n'hi ha un al clúster. Regula l'accés als fitxers per part dels clients. Manté en memòria les metadades del sistema de fitxers i controla els blocs dels fitxers que té cada DataNode.
- Els *DataNode*: Són els responsables de llegir i escriure les peticions dels clients. Els fitxers que gestionen estan formats per blocs de mida predefinida.

En aquest sentit, *HDFS* es troba optimitzat per treballar amb fitxers grans que requereixen alts fluxos de lectura i escriptura. Es altament escalable i proporciona alta disponibilitat gràcies a la replicació de blocs en diferents nodes i tolerància a les fallades.

*La capa* MapReduce permet fer consultes a una base de dades immensa i obtenir respostes ràpides. És capaç d'enviar una ordre a cada màquina del sistema distribuït perquè busqui les dades en el seu disc dur, recol·lecti totes les contestacions i les ordeni per resoldre la consulta. També s'introdueix amb més detall a l'apartat *3.4.1*

*Hadoop* MapReduce pot resoldre amb èxit càrregues de treball de gran complexitat, com el Processament del Llenguatge Natural o el *Machine Learning*. Però no és l'únic algoritme que es pot utilitzar. *Hadoop* disposa de components com *Hive, Pig, Mahout*, entre altres, que permeten utilitzar altres llenguatges per construir algoritmes de manera més àgil.

*Hadoop* compta amb conjunt de sub-projectes vinculats (alguns d'ells amb constat evolució) per proporcionar característiques addicionals, entre els quals es destaquen els següents.

- ***Spark i Storm***: motors avançats de processament de dades i que s'analitzaran en els apartats 4.2.2 i 4.2.3, respectivament.
- ***HBase***: és la base de dades distribuïda i orientada a columnes, que suporta emmagatzemament de dades estructurades per grans taules. Està construïda a sobre de *HDFS* i suporta el processament MapReduce.
- ***Mahout***: és una llibreria escalable de *Machine Learning* i *data mining*.
- ***Pig***: és una plataforma que proporciona un llenguatge d'alt nivell, anomenat *Pig Latin*, per a l'anàlisi de grans conjunts de dades i amb una infraestructura que permet l'execució de computació en paral·lel. Els programes escrits en aquest llenguatge d'alt nivell es tradueixen en seqüències de programes de MapReduce.
- ***Hive***: és una infraestructura de *Data Warehouse* per sobre de *Hadoop* que permet resum de dades, consultes *ad-hoc* i anàlisis de fitxers grans. S'utilitza *HiveQL*, un llenguatge semblant a SQL però empleat en un esquema "*on read*" (de lectura) que són automàticament convertits a treballs MapReduce.
- **Oozie**: coordinador de compoenents. Permet construir processos de negoci en els que s'encadenen execucions de *MapReduces, Pigs* i *Hives*. A més, *Oozie* també permet planificar en el temps l'execució d'aquests processos.
- ***Sqoop***: és una aplicació per a transferir dades entre bases de dades relacionals i *Hadoop*.
- ***Zookeeper***: proporciona servei de configuració centralitzada, sincronització i registre de noms.
- ***Ambari***: Una eina basada en web per a l'aprovisionament, gestió i seguiment de les agrupacions *Apache Hadoop* que inclou suport per *Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig* i *Sqoop*. *Ambari* també proporciona un panell de

control per a la visualització de l'estat del clúster com mapes de calor i la capacitat de visualitzar de forma conjunta les aplicacions MapReduce, *Pig* i *Hive* amb funcionalitats per diagnosticar-ne les característiques de rendiment d'una manera fàcil d'utilitzar.
- ***Flume***: servei per recol·lectar, afegir i moure grans conjunts de dades cap a un entorn Hadoop.
- ***Chukwa***: és un sistema de recollida de dades i monitorització per a la gestió de grans sistemes distribuïts. Emmagatzema les mètriques del sistema així com els arxius de registre en *HDFS*, i usa MapReduce per generar informes.
- ***Kafka***: sistema de missatgeria distribuït.

Altres paquets, que no solen incloure en *Hadoop* però que si s'estan fent servir en projectes Big Data són:

- **OpenNLP**: per a processament de llenguatge natural, basat en *machine learning*.
- **UIMA**: utilitzat per *IBM Watson*, és un *framework* per integrar aplicacions PLN.
- **Apache Solr**: cercador avançat, basat en *Apache Lucene*.

### *4.1.1. Les Distribucions d'Apache Hadoop*

*Hadoop* és una plataforma de codi obert, qualsevol el pot agafar, empaquetar i oferir-ho com una distribució de la plataforma. Són diverses les companyies que comercialitzen aquest tipus de solució. Aquestes distribucions inclouen moltes de les aplicacions que esmentem en el punt anterior.

A la pàgina web d'Apache podem trobar una llista completa de productes[27] que inclouen Apache Hadoop, aplicacions derivades i suport comercial.
Entre elles destacarem les següents:

- *Amazon Web Services*: coneguda com *Amazon Elastic MapReduce*
- *Cloudera*: anomenada CDH i complementada amb productes propis.
- *Hortonworks*: ofereixen una plataforma 100% *open-source* i basen la sostenibilitat de la seva empresa en els serveis. Ofereixen els seus serveis com a *partners* d'altres empreses destacades del sector, com *SAS*.
- *IBM*: anomenada *IBM InfoSphere BigInsights*
- *MapR Technologies*: distribució orientada a l'alt rendiment, han realitzat reenginyeria d'alguns dels seus components.
- *Pivotal*: ofereix la distribució *Pivotal HD*.

Altres projectes i empreses rellevants de l'entorn Big Data com *BigTop*, *Datameer* o *VMware*, són esmentats en aquesta llista, i presumiblement també aniran creixent i desenvolupant en un futur proper.

## 4.2. Motors de processament avançats

En aquest apartat s'introdueixen diferents tècniques i els nous motors de processament de Big Data. Des de la tècnica tradicional que ve utilitzant *Hadoop* a les noves tècniques que s'utilitzen en els nous motors avançats de processament que utilitzen *Apache Spark* i *Apache Storm* per fer front a l'anàlisi de dades en temps real o quasi real.

### *4.2.1. Tècniques de processament*

Existeixen diferents tècniques de processament Big Data, que van des de els clàssics a renovats sistemes de processament per lots, fins als potents sistemes de *Stream Computing* i algunes solucions mixtes.

Depenent dels requisits de latència, rendiment i tolerància a fallades es triarà un o altre sistema de processament. Serà especialment rellevant determinar si la font de dades pot, en cas de fallada del sistema de processament, disposar de nou o no d'un missatge prèviament rebut i si els missatges poden tornar-se a trobar donat un criteri de recerca, per la qual cosa serà necessari estar recolzada per una arquitectura amb sistema de fitxers distribuït.

- **Processament per lots (*Batch processing*)**: és una manera eficient de processar grans volums de dades per tractar aquelles fonts on la incorporació al sistema es faci de forma puntual o amb periodicitats mitjanes o altes. Aquest model es suficient per molts de casos (com d'indexació web). Les dades es recullen, introdueixen, processen i es produeixen els resultats per lots (*Hadoop* es centra en el processament de dades per lots). El processament per lots requereix diferents eines per a l'entrada, procés i sortida.

- **Anàlisi en temps real (*Real Time*)**: es dut a terme per els sistemes *Stream Computing* que permeten llegir i analitzar dades en temps real. Són sistemes escalables, formats per xarxes de nodes, que processen milers de missatges per segon, tolerants a fallades i fiables, que garanteixen el lliurament del missatge. Els seus models són senzills, basats en topologies, amb pocs tipus de nodes que executen tipus de tasques senzilles. Un dels sistemes més populars és *Apache Storm*.

- ***Micro-batching***: és una tècnica mixta entre els dos sistemes anteriors, permet empaquetar fluxos (*stream*) de dades entrants en paquets per al seu tractament per un sistema de processament per lots. Un exemple és *Trident*, una abstracció d'alt nivell basada en *Apache Storm*. *Trident* divideix els lots en particions, cadascuna orientada a ser executada per un node. També, un altre dels referents en Big Data, és el motor de processament de dades a gran escala *Apache Spark* a través de la seva extensió *Spark Streaming*, un sistema de computació en clústers, de propòsit general caracteritzat per la seva alta velocitat.

Una altra tendència és l'Arquitectura Lambda que intenta treure el millor dels mètodes de *stream computing* i *batch processing*. Parteix d'una font de dades base, en la qual s'emmagatzema tota la informació disponible. Disposa de tres capes: una capa "*batch*", per grans quantitats de dades; una capa anomenada "*speed*" (*stream computing*), per a fluxos de dades en temps real amb objectiu de reduir la latència i lliurant les dades en una base de dades *NoSQL*; i una capa de servidor que recull les sortides de les altres dues capes i que respon a consultes del sistema generant vistes de les dades.

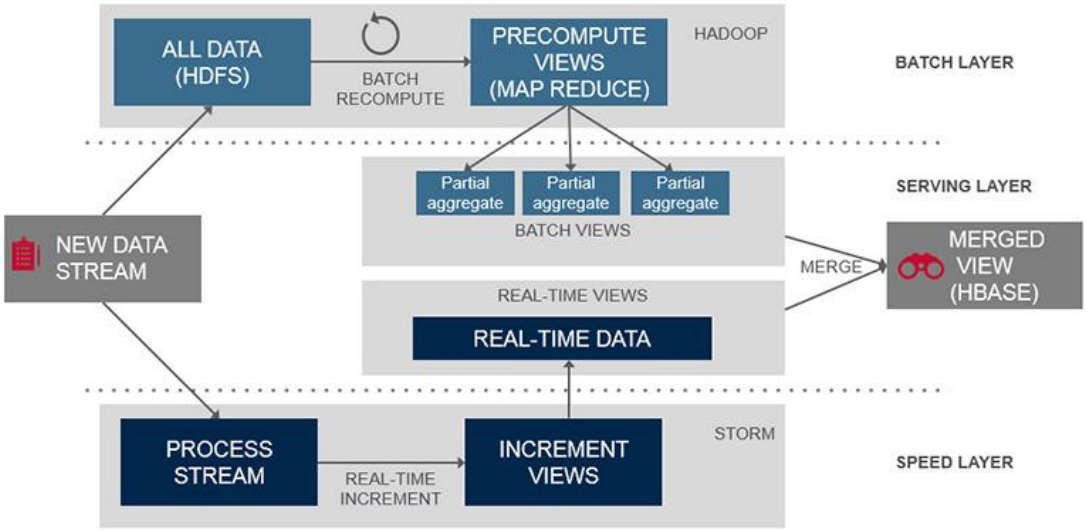


**Figura 5. Arquitectura Lambda**[28]

## *4.2.2. Apache Spark*

*Spark* sorgeix davant la necessitat de disposar d'un processament més ràpid. Com que Hadoop treballa principalment a disc, es una tecnologia molt lenta per algoritmes iteratius (el resultat de cada funció *Map* s'emmagatzema temporalment en disc, i el mateix després d'aplicar *Sort* o *Reduce*).

*Apache Spark*[29] es basa en un mòdul nucli que proporciona funcionalitat bàsica per a planificació i gestió de tasques i d'entrada i sortida de dades. Defineix un concepte especialment rellevant, denominat RDD (en anglès "*Resilient Distributed Datasets*") que constitueixen col·leccions lògiques de dades distribuïdes entre diverses màquines. *Spark* permet fer referència als RDDs a través d'*API*s. La seva arquitectura està orientada al processament amb la memòria RAM (en anglès "*in-memory*"). Això permet un rendiment molt superior en alguns tipus de processament, per exemple els que s'utilitzen en *Machine Learning*.

Per sobre dels RDDs s'executen dos tipus d'operacions, transformacions i accions. Les transformacions donen com a resultat un nou RDD, mentre que una acció obté informació del RDD. Per exemple, un RDD amb tota la informació, si es filtra per un camp s'obtindrà un nou RDD i com que les dades son inamovibles, s'aplicarà una transformació. Per altra banda, si sobre un nou RDD es consulten quants elements conte, s'aplicarà una acció.

*Spark* utilitza un mode peresós d'execució, si no es llança una acció sobre un RDD no executarà operacions MapReduce fins en aquell moment, ni portarà dades dels orígens. D'aquesta manera, en aplicar transformacions a diversos RDDs se'n generen de nous, i com que la memòria RAM és limitada, *Spark* automàticament desallotja d'aquesta els RDDs que no s'estan utilitzant. Per tal de no copsar la RAM i ha l'opció de portar a persistent les dades.

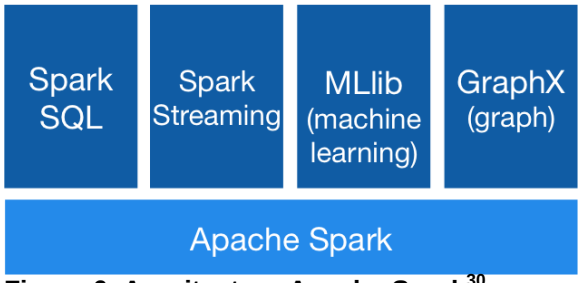


**Figura 6. Arquitectura Apache Spark**[30]

Compta amb 4 llibreries principals:

- ***Spark SQL***, que habilita les consultes mitjançant el llenguatge SQL a una abstracció anomenada *SchemaRDD*, que dóna suport a dades estructurades i semiestructurades. Té integració amb *Hive*, *JSON* i altres.
- ***Mlib*** és la llibreria de *Machine Learning* per explotar les especials capacitats de *Spark* per millorar en gran mesura el rendiment davant d'altres aplicacions. Es podem utilitzar algoritmes implementats en *Java*, *Scala* y *Python*.
- ***GraphX***, és el component (llibreria) dedicat al processament distribuït de grafs, implementant algoritmes com *Dijkstra, Primo* o *Druskal*.
- ***Spark Streaming***[31], és el component de '*near real time*' de *Spark*. A través dels RDDs definits per una finestra temporal (temps o nombre de files) permet fer anàlisis de dades en quasi temps real processant fluxos continus de dades. A les dades se'ls apliquen funcions d'alt nivell, com les conegudes *Map* i *Reduce* o funcionalitats de processament de grafs o *Machine Learning* amb *MLib* o *GraphX* i el seu resultat pot ser emmagatzemat en bases de dades, sistemes de fitxers distribuïts o publicats en sistemes de visualització Big data.

### *4.2.3. Apache Storm*

*Apache Storm*[32] és un sistema que serveix per recuperar *streams* de dades en temps real des de múltiples fonts de manera distribuïda, tolerant a fallades i en alta disponibilitat. *Storm* està principalment pensat per treballar amb dades que han de ser analitzades en temps real, com les dades de sensors que s'emeten amb una alta freqüència o dades que provinguin de les xarxes socials on a vegades és important saber què s'està compartint en aquest moment.

La topologia que implementa *Apache Storm* compta amb dos tipus de nodes, nodes "*Spouts*" i nodes "*Bolts*". Els *spouts* converteixen fluxos de dades en temps real en fluxos de tuples formades per parells clau/valor i les emeten cap

nodes *bolts* que poden executar tasques senzilles com MapReduce o accions més complexes (funcions d'un sol pas) com el filtrat, agregacions, o la comunicació (lectures/escriptures) amb entitats externes com ara una base de dades. Una topologia típica de *Storm* implementa múltiples transformacions i per tant requereix de múltiples *bolts* amb fluxos de tupla independents. Cada *spout* i *bolt* és executat en paral·lel en múltiples nodes del sistema.

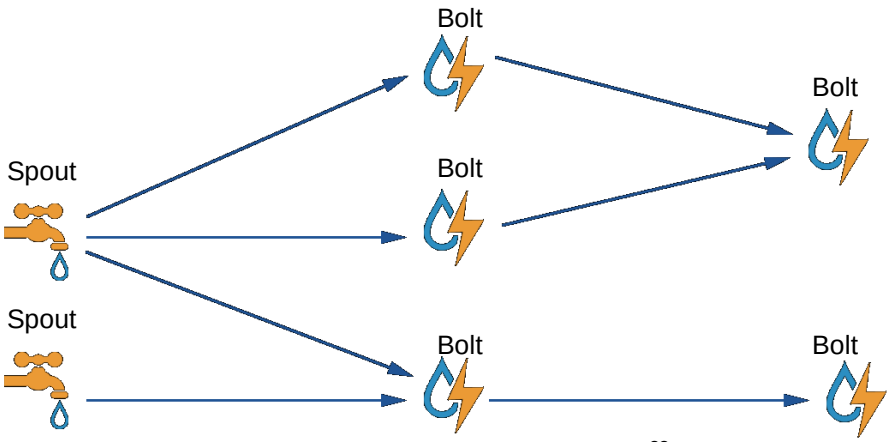


**Figura 7. Topologia Storm**[33]

Storm agrupa les tasques assegurant que totes les tuples amb els mateixos valors s 'encaminen cap a la mateixa tasca. Compta amb dues aplicacions de suport, *Nimbus* i *Zookeeper*. *Nimbus* rep la topologia dissenyada, calcula les assignacions i les envia a *Zookeeper*.

*Zookeeper* envia a nodes supervisors "*Supervisor*" les assignacions, on aquests llancen nodes treballadors "*Worker Node*" per executar la topologia. Per assegurar la tolerància a fallades, cada node treballador notifica periòdicament a *Zookeeper* que segueix actiu i aquest ho transmet a *Nimbus*. Si no arriba la notificació, el supervisor reinicialitza el node treballador. Si falla repetidament el node, sigui treballador o supervisor, *Nimbus* torna a assignar el treball a altres nodes.

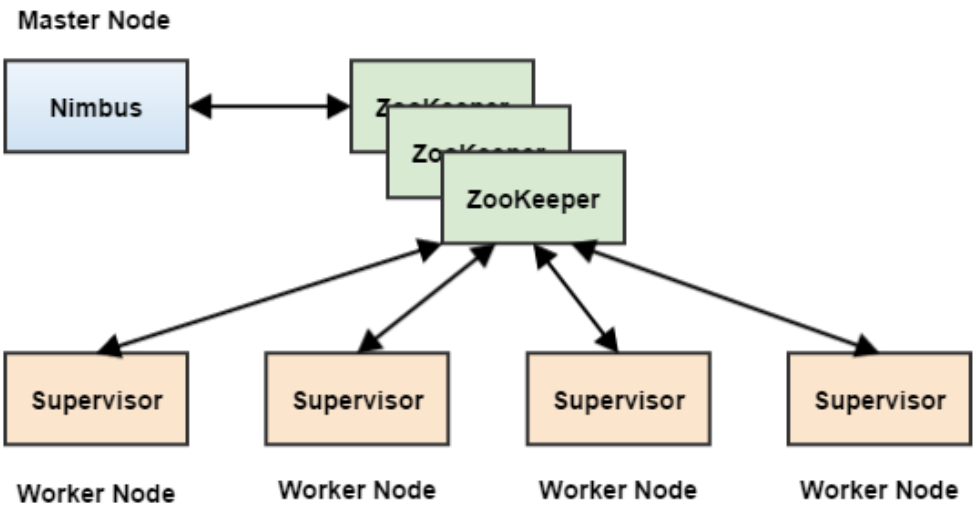


**Figura 8. Arquitectura Storm**[34]

Els nodes Bolt emeten senyals d'ACK o de FAIL per notificar que la tasca ha estat o no executada, fent així al sistema fiable. Aquests enviaments es fan a través de nodes *Bolt* especialitzats únicament en aquesta tasca.

Moltes d'aquestes tecnologies es troben disponible en distribucions que ofereixen empreses com *Cloudera*, *Datastax*, *Hortonworks* o *MapR* que permeten desplegar un clúster de forma senzilla, afegint panels d'instruments per la visualització de l'estat del clúster i oferint suport als clients.

## 4.3. Bases de dades NoSQL

Les bases de dades *NoSQL, que significa 'Not Only SQL'* (*No tan sols SQL*) com una manera de destacar que existeixen tipus de dades diferents a les relacionals, són un conjunt de Sistemes Gestors de Bases de Dades que foren desenvolupades o promogudes en mode *open-source* per empreses com *Google*, *Amazon* i altres davant les necessitats del seu model de negoci i a partir de llavors han seguit evolucionant.

Els seus trets generals, ja s'han introduït en l'apartat '3.3.1 Bases de Dades *Big Data*'. Existeixen varies classificacions per a les bases de dades *NoSQL*, sent les més importants les següents[35].

- Clau - Valor: a partir d'una clau es recupera un objecte binari. *DynamoDB, Redis, BerkeleyDB, GenieDB, Voldemort, Oracle NoSQL* i *Windows Azure NoSQL* serien les més destacades.

- Big Table / columnars: sistemes que reparteixen les files i les columnes d'una taula en diferents servidors. D'aquest tipus són les molt conegudes *Apache HBase, Cassandra, Hypertable* o *Amazon SimpleBD*.

- Documents: manegen conjunts de dades identificades per etiquetes, usat habitualment per emmagatzemar informació de formularis emplenats pels usuaris. Exemples d'aquest tipus serien *MongoDB, Elasticsearch, CouchBase o CouchDB*. Dins d'aquest grup es separen les de Documents XML, bases de dades especialitzades en gestionar documents en format XML com *BerkeleyDB XML, EMC Documentum* o *BaseX*.

- Grafs: representació d'informació estructurada en forma de xarxa. *Neo4J, InfiniteGraph* i *OpenLink Virtuós* serien els productes més significatius.

També es pot considerar un cinquè tipus de base de dades *NoSQL*, els índexs *'fulltext'*, que són textos no estructurats i són la base dels coneguts *Apache Lucene* i *Apache Solr*.

Hi ha altres grups, com les bases de dades orientades a objectes, solucions *grid & cloud*, bases de dades multidimensionals, orientades a esdeveniments o a xarxes, a més de molts altres tipus orientades a propòsits específics.

A continuació es descriuen els 4 grups de bases de dades *NoSQL* que destaquen més en el mercat: clau/valor, orientades a columnes, orientades a documents i orientades a grafs.

### *4.3.1. Bases de dades clau/valor.*

Aquest tipus de bases de dades destaquen per la seva alta escalabilitat. Abasten bé projectes amb textos estructurats i semiestructurats, dades de xarxes socials, logs de servidors web i la majoria de les dades orientades a negoci, de manera que aquest tipus de bases de dades són de les més utilitzades en projectes Big Data. També són utilitzades quan es realitzen grans volums d'escriptures en múltiples nodes o quan es realitzen analítiques a gran escala a grans clústers.

Encaixen molt bé en projectes en què l'escriptura es realitza una única vegada i es realitzen moltes lectures. En conseqüència, necessiten emmagatzemar la informació i recuperar-la a alta velocitat. Són utilitzades per exemple en situacions com la de gestió de les sessions d'usuari, caracteritzades per accés ràpid a lectures i escriptures i per no necessitar durabilitat de les dades. Un altre exemple molt clar és la participació en una xarxa social, (*Facebook*, *Twitter* ...) que és escrita una única vegada per l'usuari autor i és llegida a continuació pels seus seguidors, amics o el concepte que estigui implementat a la xarxa social.

### *4.3.2. Bases de dades de famílies de columnes*

Són molt utilitzades en entorns analítics com OLAP (*On Line Analytic Processing*), els coneguts "cubs multidimensionals" tan utilitzats en entorns financers i de màrqueting), tradicionals en el món del Business Intelligence, des del qual ha evolucionat el Big Data.

Aquest tipus de bases de dades estan orientades a les columnes de dades, en lloc de als registres com les bases de dades relacionals. Quan escrivim el registre d'una transacció, estem escrivint els valors de tots els camps: nom del comprador, import de la transacció, nom del venedor, el producte venut, el nombre d'unitats, etc. Les bases de dades columnars estarien orientades a gestionar de cop tots els valors dels compradors, que estarien en la mateixa columna, d'aquí el nom. De fet l'origen d'algunes bases de dades columnars és la necessitat d'emmagatzematge de columnes de dades.

Són molt utilitzades quan s'executen treballs tipus MapReduce, quan cal actualitzar i emmagatzemar registres únics, com ara tot l'històric de relació amb un client. També són bones executant el càlcul de mètriques d'una columna o un conjunt de columnes. En canvi, si han d'analitzar o escriure files, és a dir, registres, el seu rendiment no és bo.

Un dels seus precursors més rellevants és la base de dades *Google Big Table*, per aquest motiu a aquest tipus de bases de dades també se'ls anomena '*Big Table'*, a més de "Orientades a Columnes".

### *4.3.3. Bases de dades documentals*

En les Bases de dades Documentals cada document es tractat com un únic registre. Gestionen molt bé text no estructurat i particularment bé text semiestructurat, és a dir, text codificat segons un esquema conegut, com XML, JSON, YAML, PDF, correu electrònic o fins i tot documents ofimàtics.

Són molt bones per tant en recuperació de coneixement o temes inclosos en grans conjunts d'informes i documentació o recerca de correus electrònics. Hi pot ser facilitada afegint metadades, claus i llenguatges específics dependents del model de base de dades utilitzat.

Concretament són molt utilitzades per a recerca de patents, recerca de precedents legals, recerca de papers científics i dades experimentals. Així mateix són molt bones per a integrar diferents fonts de dades que poden residir en tipus de bases de dades incompatibles.

Els programadors les consideren bases de dades amigables, que permeten un modelatge de dades natural i desenvolupament ràpid. A més encaixen molt bé amb el paradigma programació orientada a objectes, possiblement el paradigma dominant en l'actualitat.

### *4.3.4. Bases de dades orientades a grafs*

Aquest tipus de bases de dades estan inspirades pels treballs de *Leonhard Euler* i la teoria de grafs. Són especialment útils quan les dades estan molt interconnectades i no són tabulars. S'utilitzen en tot tipus d'aplicacions relacionades amb la web semàntica, amb ontologies, i són també molt utilitzats en emmagatzematge d'imatges i quan en les dades estan implicades algoritmes sustentats en la teoria de grafs. Enllacen ràpidament persones, productes, compres i qualificacions, per exemple.

Han de permetre executar com a transacció única qualsevol consulta que exploti les relacions entre entitats. En una base de dades relacional, per buscar relacions entre entitats de negoci relacionades, hauríem d'anar pas a pas, relació a relació, executant recerques. En una base de dades orientada a grafs això s'executaria en una única transacció. S'utilitzen llenguatges especialment dissenyats, com SPARQL.

Es poden fer servir com una base de dades de propòsit general però requereix un canvi de paradigma a l'hora de dissenyar les relacions entre les dades ja que només ofereixen bon rendiment amb dades molt interconnectades.

# 5. Big data en Sanitat

El sector sanitari des de que es va a començar a digitalitzar la Història Clínica ha experimentat un creixement substancial en l'acumulació de grans volums de dades provinents de seus sistemes. En primer lloc dels sistemes d'Història Clínica Electrònica (HCE) i registres mèdics electrònics (EMR); posteriorment, radiografies, ressonàncies magnètiques en 3D, mamografies, TACs en 3D, altres imatges i dades generades per una amplia varietat aparells de diferents serveis mèdics, digitalització d'històric i de documentació en paper, etc.

Totes aquestes dades que genera l'àmbit sanitari també es poden trobar amb en diferents formats de varietat.

- Dades estructurades: en aquest cas es corresponen a bona part de les dades dels sistemes HCE com són les d'admissió i hospitalització, dades demogràfiques, una part dels registres mèdics, dades de triatge, codificacions de malalties i procediments, lectures d'instruments de laboratori, sensors que recullen constants vitals, i altres aparells biomètrics que disposen d'un alt nivell d'integració. També les dels sistemes ERP com dades de comptabilitat, facturació, recursos humans, compres, etc.

- Dades no estructurades: documents PDF que poden ser generats per digitalització d'històric en paper o altres sistemes mèdics poc integrats; imatges per el diagnòstic en diferents formats JPG, TIF, DICOM, MPG, entre altres; correus electrònics i documents de text.

- Dades semiestructurades: Els registres mèdics poden seguir una certa estructuració, però una part és de text lliure o imatge. Un exemple són els l'HTML o l'XML.

En el sector sanitari, les dades no estructurades representen el 80% del total de les dades de salut, tot i que quasi la seva totalitat prové de les imatges per el diagnòstic.

Més recentment, i cada cop amb major quantitat, es podrien anar afegint nous fluxos de dades provinents de fonts externes, principalment generats per aparells de fitness, la genètica i la genòmica, mitjans de comunicació social, la investigació mèdica. Tot i que, en l'actualitat poques d'aquestes dades es poden capturar, emmagatzemar i organitzar de manera que puguin ser tractades i analitzades per extreure informació útil.

Tal i com es desprèn de la figura 9, la sanitat es un dels sectors que genera grans quantitat de dades. Existeixen 16.000 hospitals arreu del món que recopilen informació dels seus pacients i 4,9 milions de persones disposen d'algun tipus de dispositiu que monitoritza la seva salut. Aquesta infografia[36] ha estat realitzada per l'àrea de salut de Siemens a través d'estudis de consultores i empreses especialitzades. També recull 6 formes en que Big Data pot transformar la sanitat.

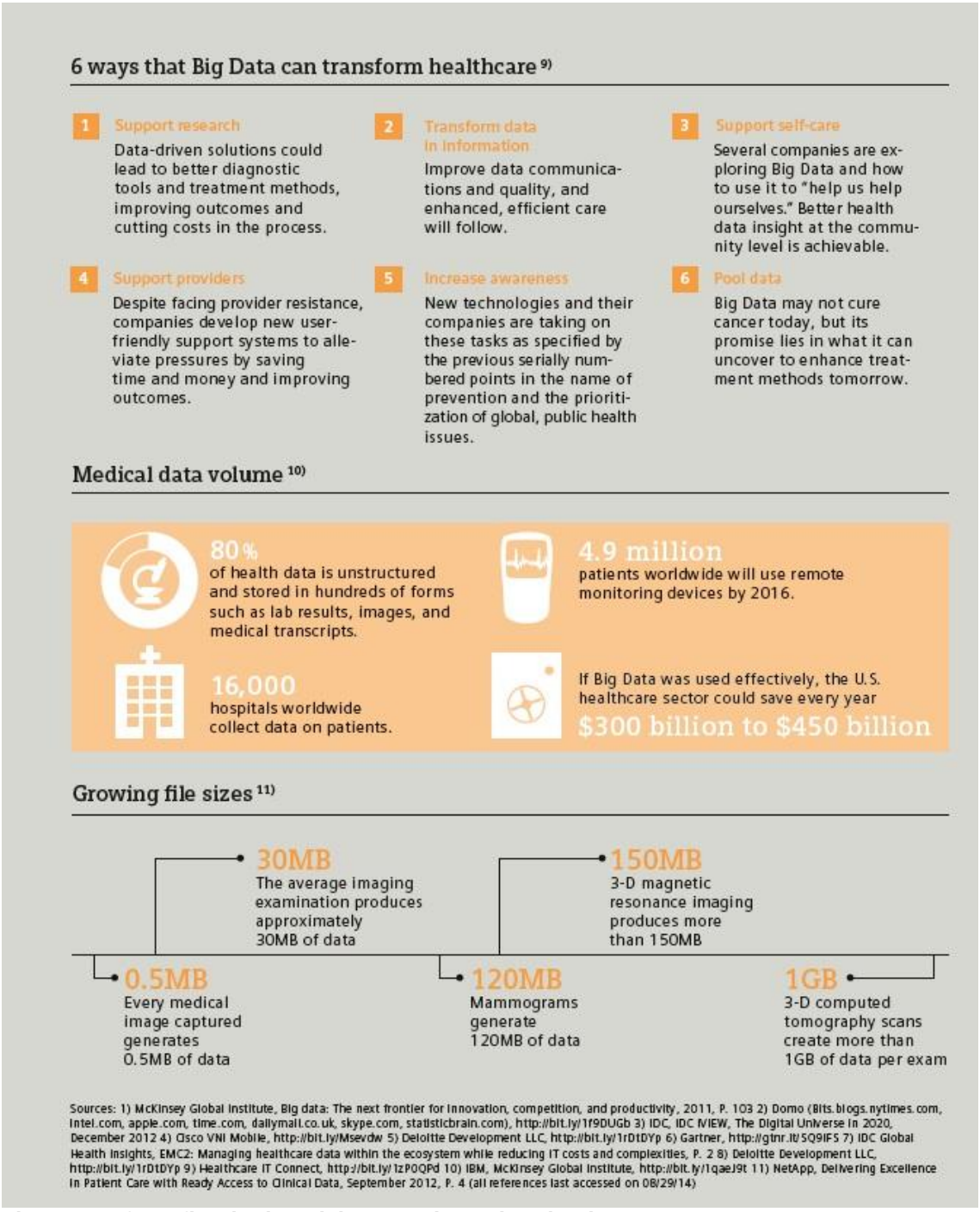


**Figura 9. Infografia Big data del sector de sanitat de Siemens**

El potencial de grans volums de dades en l'assistència sanitària rau en la combinació de dades tradicionals amb les noves formes de dades[37], tant individual com a nivell de població. Amb l'aparició de les tecnologies que permeten l'anàlisi de grans volums de dades ja s'està veient com els conjunts de dades a partir d'una multitud de fonts pot donar suport a la investigació i al descobriment més ràpid i fiable.

## 5.1. Promeses i potencial

McKinsey en el seu estudi "*The next frontier for innovation, competition, and productivity*"[9] creu que Big Data podria ajudar a reduir els costos i ineficiències en les àrees d'Operativa clínica, d'Investigació i Desenvolupament i en la Sanitat Pública.

A més, segons Raghupathi *en el seu article "Big data analytics in healthcare: promise and potential*"[38] apunta que l'analítica de dades massives en l'assistència sanitària pot contribuir a:

- La medicina basada en evidència: Combinar i analitzar una varietat de dades estructurades i no estructurades de registres EMR, financeres i operatives, dades clíniques i les dades genòmiques per a que coincideixin els tractaments amb resultats, predir els pacients en risc de contraure una malaltia o el reingrés hospitalari i proporcionar una atenció més eficient.

- L'anàlisi de genòmica: realitzar la seqüenciació de gens de manera més eficient i rendible i fer l'anàlisi genòmica una part del procés de decisió en l'atenció mèdica normal i la creixent història clínica del pacient.

- L'anàlisi de fraus: analitzar ràpidament un gran nombre de sol·licituds de reclamació per reduir el frau, les ineficiències i l'abús.

- Integració de dispositius i monitorització remota: Capturar i analitzar en temps real els grans volums de dades en ràpid moviment tant intrahospitalària com a la llar amb dispositius portables, per a la supervisió de la salut i la predicció d'esdeveniments adversos.

- L'anàlisi dels perfils de pacients: Aplicar anàlisi avançats per perfils dels pacients (per exemple, la segmentació i modelatge predictiu) per identificar les persones que es beneficiarien de canvis d'atenció o d'estil de vida proactives, per exemple, aquells pacients amb risc de desenvolupar una malaltia específica (per exemple, diabetis) que el faria beneficiar-se de l'atenció preventiva.

## 5.2. Barreres

Tot aquest potencial no esta lliure de problemàtiques, segons Garcia, J. en la seva publicació "*La medicina del futuro pasa por Big Data*" [39] i coautor de la publicació de Telefònica "*Big data. El poder de convertir datos en decisiones*"[40]: "Per poder treure el màxim partit a les tecnologies de Big Data en la sanitat seria necessari capturar, emmagatzemar i analitzar totes les dades disponibles sobre assajos clínics, historials mèdics, seqüenciació d'ADN de pacients, informació procedent de xarxes socials... S'hauria de disposar, per tant, d'una enorme base de dades compartida entre tots els hospitals i resta d'agents del sector de la salut".

L'autor agrupa en tres les barreres existents:

- **Barrera administrativa**: cal que hi hagi un acord entre totes les parts involucrades per dur a terme la compartició d'informació que, a dia d'avui, resideix en compartiments estancs. Això no serà trivial si tenim en compte que a Espanya són les comunitats autònomes les que han assumit les competències en matèria de sanitat pública. I la situació es complica encara més si es pretén involucrar en un marc de col·laboració comú també a companyies asseguradores, farmacèutiques i hospitals privats.

- **Barrera tecnològica**: La tecnologia inclosa en els projectes de Big Data ja és una realitat fa algun temps, i els seus pilars fonamentals són els sistemes d'arxius distribuïts, bases de dades escalables, programari de tractament massiu (tipus *Hadoop*), *cloud computing* i Internet de les coses. Però aquesta tecnologia ha de consolidar-se encara en el sector sanitari, pel que serà necessari que augmentin les inversions públiques i privades en aquest tipus de solucions. A més, potser el factor més important sigui l'humà, és a dir, els científics de dades. És crucial comptar amb la presència d'analistes de dades experts en l'àmbit de la salut perquè, a través de l'ús de tecnologies Big Data , puguin donar el suport adequat als metges en la presa de decisions relatives als seus pacients.

- **Barrera legal**: Perquè Big Data pugui entrar en escena i s'aconsegueixin els millors resultats, cal emmagatzemar una ingent quantitat de dades, procedents majoritàriament de pacients. Aquestes dades personals són extremadament sensibles i cal garantir el compliment de la LOPD (Llei Orgànica de Protecció de Dades) per assegurar la seva confidencialitat i integritat. Aquesta barrera serà fàcilment salvable si es compta en tot moment amb l'ajuda d'experts en seguretat de la informació.

## 5.3. La privacitat de les dades personals

La privacitat en les dades de salut és un aspecte que agafa especial atenció en parlar de projectes Big Data ja que s'està convertint en tot un repte, tant per a les administracions públiques com per a les empreses.

Els prestadors sanitaris, públics i privats, porten tres dècades informatitzant la prestació sanitària i recollint, de mica en mica, dades dels ciutadans com a pacients. Però només recentment aquests sistemes -fins fa poc illes d'informació- han començat a interconnectar-se i a crear grans conglomerats de dades. Que ha d'estar protegida contra accessos no autoritzats.

La privacitat de les dades de salut dels ciutadans és un dret generalment reconegut en gairebé tots els països del món. Tal i com introdueix Sánchez, JJ. de Telefònica en el seu article "*La privacidad de los datos de salud en la era digital (2015)*"[41] i continua: A la Unió Europea, la privacitat de les dades dels ciutadans, incloses les de salut, es recull en la directiva 95/46/EC, que es trasllada a la legislació nacional espanyola en la Llei Orgànica de Protecció de Dades (LOPD).

La LOPD estableix tres nivells de protecció de les dades personals, i les dades de salut gaudeixen del nivell de protecció màxim. Entre les mesures que administracions públiques i empreses han de complir es troben:

- Requisits relacionats amb la identificació en l'accés a la informació confidencial.
- Requisits relacionats amb la gestió de documents i mitjans electrònics d'emmagatzematge.
- Requisits relacionats amb les còpies de seguretat, el seu emmagatzematge i restauració.
- Requisits relacionats amb les comunicacions i el seu xifrat.
- Requisits relacionats amb l'accés físic de les persones a les instal·lacions i suports físics de la informació confidencial.
- Requisits relacionats amb la gestió d'incidents de seguretat.
- Requisits relacionats amb la devolució de dades als seus propietaris.

Als Estats Units, la privacitat de les dades de salut es recull en l'anomenada *HIPAA*, això és, la *Health Insurance Portability and Accountability Act*, que estableix:

- Una normativa de privacitat (*Privacy Rule*), orientada a regular l'ús que es fa per part de les organitzacions sanitàries de la informació confidencial dels pacients.
- Una normativa de seguretat (*Security Rule*), que estableix les normes de seguretat informàtica que han de posar en marxa els prestadors sanitaris que fan servir informació de pacients en format electrònic (*electronic protected health information* o *e-PHI*). Entre les exigències d'aquesta última norma hi ha el control i auditoria d'accessos a informació confidencial, el xifrat de les comunicacions, el control d'accés físic als centres de procés de dades on es guarda informació confidencial, la configuració i ús dels llocs de treball des dels quals es pot accedir a informació confidencial, procediments d'accés a la informació, formació del personal i avaluació de riscos, entre d'altres.

La *HIPAA* és d'obligat compliment per a totes les asseguradores i empreses associades que manegen dades confidencials de pacients, i l'Oficina de Drets Civils (*Office of Civil Rights, OCR*) persegueix el seu compliment. Però l'escenari de la seguretat de les dades de salut és susceptible encara de complicar-se i ho està fent. El nou repte ve de la mà de les *apps* en l'entorn sanitari. Des de fa poc, estan apareixent al mercat nombroses aplicacions orientades al que en anglès s'anomena *self tracking* o autoseguiment, amb les que els individus poden registrar les seves dades de salut i les constants vitals que constitueixen la seva evolució.

Aquestes aplicacions, a vegades aïllades i altres vegades venudes amb dispositius que capturen les constants vitals de l'usuari (tensió, pes, saturació d'oxigen en sang, activitat física ...) deixen la informació en mans de les empreses fabricants i aquesta acaba normalment en magatzems de dades en

el núvol als EUA, fora del control d'una cosa tan arcaic i necessari com les legislacions nacionals.

Els gegants del sector, com Apple i Google, en veure el potencial comercial d'aquestes dades, han generat una nova generació de "agregadors" de dades de salut: *Health Kit* i *Google Fit*. I han començat una carrera per connectar tant dispositius com a grups hospitalaris, per ara als EUA, el que suposa, a la fi, poder pujar a les seves respectives núvols les dades de salut de milions d'usuaris.

Sense una anàlisi jurídica en profunditat de la matèria, sembla evident que aquestes aplicacions estan volant per sota del radar de la LOPD i la resta de translacions nacionals de la directiva europea de protecció de dades. Fins i tot als EUA la *HIPAA* és d'obligat compliment per a les asseguradores sanitàries i els seus associats, de manera que les empreses relacionades amb el *self tracking* ni tan estan obligades a complir la mateixa (la qual cosa s'ha convertit en una de les principals crítiques a aquesta legislació).

## 5.4. Arquitectura de Big Data en Sanitat

Per tal d'introduir a com es pot articular un sistema Big Data Analitys en l'entorn sanitari i aprofundir en els beneficis potencials que hi pot aportar, cal entendre la seva arquitectura[42] i la funcionalitat seus components.

L'arquitectura de Big Data consta d'un marc lògic basat en les dades que s'inicia amb la captura de dades, procedeix a través de la transformació de dades, i acaba amb el consum de dades. La figura 10 mostra una de les millors pràctiques d'arquitectura Big Data que es composta de cinc capes arquitectòniques principals: (1) les dades, (2) l'agregació de dades, (3) l'analítica, (4) la recerca d'informació, i (5) el govern de Big Data.

Aquestes capes lògiques constitueixen els components de Big Data que realitzen funcions específiques, i per tant permetrà entendre com transformar les dades de sanitat de diverses fonts en informació clínica útil a través implementacions Big Data.

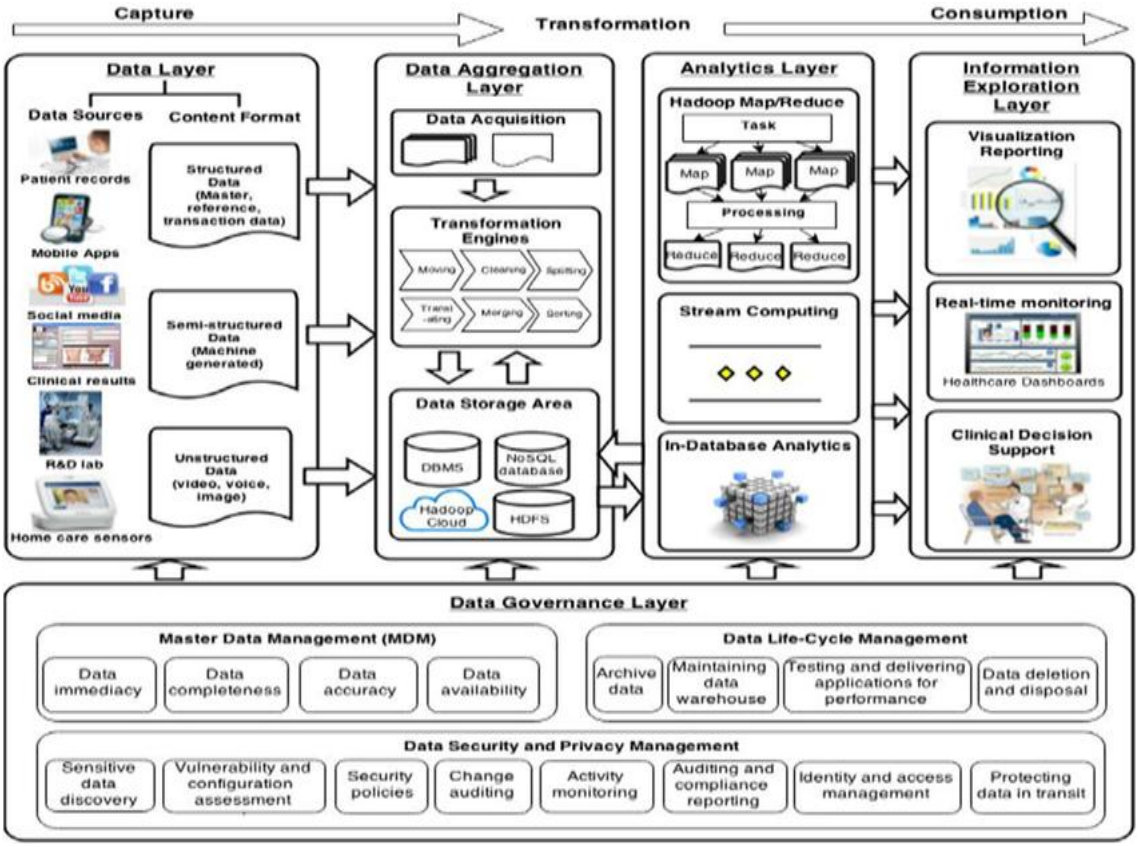


**Figura 10. Model d'Arquitectura Big Data en sanitat**

### *5.4.1. Capa de dades*

Aquesta capa inclou totes les fonts de dades disponibles per proporcionar els coneixements necessaris per a donar suport a les operacions diàries i resoldre problemes de negocis, ja sigui per ser una part intrínseca i necessària del projecte, o bé amb l'objectiu d'enriquir les dades i obtenir , en conseqüència, solucions als casos d'us i necessitats de negoci de més qualitat.

Les dades podran ser estructurades, semiestructurades i no estructurades. Aquestes dades es recullen de diferents llocs dins de l'hospital o fonts externes, i seran emmagatzemades immediatament en bases de dades apropiades, depenent del seu format. Per aquest fet, les taxonomies que classifiquen aquestes fonts són rellevants.

### *5.4.2. Capa d'agregació de dades*

Aquesta capa és responsable de gestionar les dades de les diferents fonts de dades. En aquesta capa, les dades seran digerides de manera intel·ligent mitjançant la realització de tres passos: adquisició de dades, transformació, i emmagatzemament.

L'objectiu principal d'adquisició de dades és llegir les dades proporcionades per les diverses fonts, freqüències, mides i formats. Aquest pas pot ser un obstacle important en les primeres etapes de la implementació de Big Data, pel fet que

aquestes característiques de les dades entrants poden variar considerablement. Per tant, cal avaluar la capacitat de l'emmagatzemament , tant amb la mida com en evitar colls d'ampolla per la carrega de treball i el cost que se'n deriva.

Durant l'etapa de transformació, el motor de la transformació ha de poder moure, netejar, dividir, traduir, fusionar, classificar, i validar les dades. Per exemple, les dades estructurades que normalment contenen els registres mèdics electrònics podrien ser extrets dels sistemes HCE i, posteriorment convertir-les en un format específic de dades estàndard, segons el criteri especificat (per exemple, el nom del pacient, la ubicació o els processos clínics), i després validar el registre amb normes de qualitat de dades.

Finalment, les dades son carregades a les bases de dades de destí, com ara els sistemes de fitxers distribuïts de *Hadoop* (HDFS) o en un núvol *Hadoop* per al seu posterior processament i anàlisi. Els principis d'emmagatzematge de dades es basen en la regulació de conformitat, les polítiques de govern de dades i el controls d'accés. Els mètodes d'emmagatzematge de dades poden ser implementats en processos per lots o en temps real.

### *5.4.3. Capa d'anàlisi*

Aquesta capa és responsable de processar tot tipus de dades i de realitzar els anàlisis apropiats. En aquesta capa, l'anàlisi de dades es pot dividir en tres components principals*: Hadoop Map/Reduce*, *Stream Computing*, i anàlisi dins la base de dades (*In-Database Analitycs*), en funció del tipus de dades i el propòsit de l'anàlisi.

MapReduce és el model de programació més utilitzat en anàlisi de grans volums de dades i proporciona la capacitat de processar grans volums de dades en forma de lots de manera eficient, a més de permetre l'anàlisi de dades tant estructurades com no estructurades en un entorn de processament paral·lel massiu (MPP) .

*Stream Computing* pot donar suport al processament d'alt rendiment per un flux constant de dades en temps real o proper al real. Amb una anàlisi en temps real, els usuaris poden realitzar un seguiment de dades en moviment, respondre a esdeveniments inesperats a mesura que ocorren i ràpidament determinar les millors les accions. Per exemple, en el cas de detecció de frau en l'atenció mèdica, *stream computing* és una important eina analítica que ajuda a predir la probabilitat de les transaccions il·legals o mal ús deliberat dels comptes de clients. Les transaccions i comptes seran analitzats en temps real que permetrà generar alarmes immediatament per prevenir un gran nombre de fraus en tots els sectors de la salut.

*In-database Analitycs* es refereix a un mètode de mineria de dades construït sobre una plataforma analítica que permet que les dades siguin processades dins un *Data Warehouse*. Aquest component proporciona alta velocitat de processament en paral·lel, escalabilitat i característiques d'optimització orientats a Big Data Analytics, i ofereix un entorn segur per a la informació

empresarial confidencial. No obstant això, els resultats previstos d'anàlisi *in-database* no són en temps real i per tant és probable que generi els informes amb una predicció estàtica. En general, aquest component analític en les organitzacions sanitàries és útil per donar suport a la pràctica de la medicina preventiva i la millora de la gestió farmacèutica.

### *5.4.4. Capa d'exploració de la informació*

Aquesta és la que genera les sortides tals com les diverses opcions de visualització d'informació, monitoratge en temps real de la informació i el coneixement significatiu del negoci, que deriven de les plataformes d'anàlisi i es presenten als usuaris de l'organització. Similars a les plataformes d'intel·ligència de negoci tradicionals, la presentació d'informes és una funció crítica de Big Data que permet que les dades es puguin visualitzar de forma útil per donar suport a les operacions diàries dels usuaris i ajudar als directius a prendre decisions més ràpides i millors.

Tot i així, la sortida més important per l'atenció sanitària pot, molt bé ser, la visualització de la informació en temps real, com ara alertes i notificacions proactives, la navegació de dades en temps real, i els indicadors clau de rendiment (KPI).

Aquesta informació és analitzada a partir de fonts com ara els telèfons intel·ligents i dispositius mèdics personals. D'aquesta manera, es poden enviar als usuaris interessats o posar-se a disposició en forma de quadres de comandament en temps real per al seguiment de la salut dels pacients i la prevenció d'esdeveniments mèdics accidentals.

La capa d'anàlisi també proporciona suport excepcional per a les pràctiques mèdiques basades en l'evidència. Això ho pot fer mitjançant l'anàlisi dels registres de salut electrònics, els patrons d'atenció clínica, l'experiència mèdica i els hàbits individuals dels pacients juntament amb el seu historial clínic.

### *5.4.5. Capa de govern de dades*

El govern de Big Data és el pilar de d'aquesta arquitectura, ja que afecta a totes les capes lògiques. Aquesta capa està composta per la gestió de les dades mestres (MDM), la gestió del cicle de vida de les dades, i la gestió de la privacitat i seguretat de les dades, posant especial atenció en com aprofitar les dades de l'organització.

El primer component d'aquesta capa, gestió de dades mestres, es consideren els processos, la governabilitat, les polítiques, els estàndards i les eines per a la gestió de les dades. Les dades són estandarditzades, netejades, i incorporades amb la finalitat de crear la immediatesa, integritat, exactitud i disponibilitat de les dades mestres per donar suport l'anàlisi de dades i presa de decisions.

El segon component, la gestió del cicle de vida de les dades, és el procés de gestió de la informació de negocis al llarg de tot el seu cicle de vida, des de l'arxivat de dades, mitjançant el manteniment magatzem de dades, provant i lliurant diversos sistemes aplicables per eliminar i disposar de dades.

Mitjançant la gestió de dades de manera efectiva durant la seva vida útil, les empreses estan més ben equipades per proporcionar ofertes competitives per satisfer les necessitats del mercat i proporcionar suport als objectius de negoci amb menor temps i cost.

El tercer component, és la gestió de la seguretat i privacitat de les dades, és la plataforma per facilitar l'activitat de les dades a nivell empresarial, en termes de descobriment, avaluació de la configuració, supervisió, auditoria, i la protecció.

És essencial aplicar polítiques rigoroses i mecanismes de control per la protecció de les dades sanitàries, que són altament sensibles, per tal d'evitar violacions de la seguretat i protegir la privacitat del pacient. Mitjançant l'adopció de les polítiques de seguretat s'aconsegueix evitar els accessos no autoritzats i s'assegurarà que el nou sistema compleix amb els reglaments sanitaris. Així, es crea un ambient segur per a l'ús adequat de la informació del pacient.

# 6. El Cloud: facilitador de Big Data Analytics

La recerca de noves oportunitats, la innovació i experimentació en àrees diferents i l'esforç per replantejar els processos i models de negoci per continuar millorant són alguns dels objectius que donen suport als projectes Big Data de moltes empreses. Les organitzacions volen generar valor a partir de les seves dades i evolucionar, així com el ritme del mercat, les empeny a fer-ho. Per contra, les organitzacions estan trobant moltes dificultats per aplicar eficaçment la tecnologia Big Data a causa que s'enfronten a costos significatius en termes de l'adquisició de la infraestructura i l'obtenció de mà d'obra especialitzada.

En aquest sentit, a part de l'anàlisi de grans volums de dades, una altra iniciativa de les Tecnologies de la Informació (TI), que també tenen present moltes organitzacions de tot el món, és el *Cloud Computing*, que es tradueix com 'computació en el núvol'. La computació en núvol, vist com un model de prestació de serveis de TI, té el potencial per millorar l'agilitat del negoci i la productivitat alhora que permet una major eficiència i reduir els costos.

La computació en núvol s'està convertint en una realitat per a moltes empreses, generalment amb implementacions privades. Aquesta tecnologia està madurant i abordant les barreres en quant a l'adopció de millores en la seguretat i la integració de dades, mentre que els proveïdors de TI estan evolucionant per donar suport a la prestació de serveis en el núvol. Com a resultat, les empreses estan demostrant cada vegada més confiança en els models de prestació en núvol per implementar una infraestructura que lis permeti la transformació del teixit empresarial a mesura que van creixent.

Segons un estudi de mercat d'IDC[43] sobre la infraestructura de centres de dades de Big Data, es preveu que a la regió EMEA (Europa, Orient Mitjà i Àfrica) la infraestructura arribi a triplicar-se per a l'any 2019 amb un total de 5.400 bilions de dòlars d'inversió que es destinaran a:

- Emmagatzematge: es preveu que aquesta capacitat arribi a 20 exabytes en només tres anys.

- Recursos del núvol orientats al *Analytics*: IDC preveu que la càrrega de treball en el núvol públic augmenti del 13% de 2015 al 34% el 2019. A causa d'això, la capacitat d'emmagatzematge del núvol s'incrementarà considerablement, creixent del 25% actual al 55%. Un augment més que significatiu, tot i que la majoria de les empreses no mouran totes les seves dades al núvol sinó que es decidiran per la implementació de solucions híbrides per conservar les dades crítiques i informació més sensible en entorns locals.

- En recursos de maquinari: que passaran d'un 6% el 2015 a un 16% en 2019.

## 6.1. Models de desplegament Cloud computing

La computació en núvol és un model per habilitar l'accés a la xarxa a un conjunt compartit de recursos informàtics configurables[44] (en general, xarxes, servidors, emmagatzematge, aplicacions i serveis) de manera convenient i sota demanda, que poden ser ràpidament aprovisionats i alliberats amb el mínim esforç d'administració o interacció amb el proveïdor d'aquests serveis. Segons el seu desplegament es solen classificar generalment en tres models.

- **Núvol privat** (*Private cloud*). La infraestructura del núvol està preparada per a l'ús exclusiu d'una sola organització que comprèn diversos consumidors (per exemple, unitats de negoci). Es pot disposar d'aquest en propietat, administrat i operat per l'organització, un tercer, o alguna combinació d'ells, i pot estar *on-premises* o no.

- **Núvol públic** (*Public cloud*). La infraestructura del núvol està preparada per a l'ús obert del públic en general. Pot ser de propietat, ser administrat i operat per una organització empresarial, acadèmica, o governamental, o alguna combinació d'ells. La infraestructura està les instal·lacions del proveïdor del núvol.

- **Núvol híbrid** (*Hybrid cloud*). La infraestructura al núvol és una composició de dues infraestructures de núvol privat i públic que romanen entitats úniques, però estan unides per la tecnologia estandarditzada o propietària que permet la portabilitat de les dades i aplicacions (per exemple, equilibri de carrega en de pics de demanda de computació).

## 6.2. Cloud i Big Data una combinació apropiada

Els models de prestació de núvol ofereixen una flexibilitat excepcional, permetent als departaments TI avaluar la millor estratègia per les necessitats de cada usuari de negocis. Per exemple, les organitzacions que ja disposen d'un núvol privat poden afegir l'anàlisi de grans quantitats de dades utilitzant un proveïdor de serveis en el núvol, o construir un núvol híbrid que preserva les dades sensibles en el núvol privat, i aprofitar-se de noves fonts de dades i aplicacions externes desplegades en els núvols públics.

Segons l'informe d'IDC, "*Construir una Cloud híbrida: La TI como servicio*"[45], és precisament el nostre país el que ha aconseguit minimitzar més els costos gràcies al núvol, en comparació amb altres nacions europees. La mitjana de l'estalvi se situa al voltant del 15%, tot i que les empreses que han aconseguit implementacions amb més eficàcia han aconseguit augmentar aquesta xifra més enllà del 50%.

IDC pronostica que, de les empreses de implementaran algun model de núvol en el pròxim any, el 80% ho faran en un model híbrid, tot i que, actualment, més del 60% de les empreses no estan encara del tot preparades per fer el salt, ja que no compten amb les capacitats suficients per dissenyar i / o executar estratègies en aquesta línia.

Els models de computació en el núvol poden ajudar a accelerar el potencial de les solucions d'anàlisi escalables tal com inica l'estudi d'Intel "*Big Data in the Cloud: Converging Technologies*"[46].Les organitzacions que utilitzen la infraestructura de núvol per proporcionar analítica de grans volums de dades tenen múltiples opcions. Per factors de càrrega de treball, el cost, la seguretat i la interoperabilitat de les dades, els departaments TI poden optar per: utilitzar el seu núvol privat per mitigar el risc i mantenir el control; o utilitzar infraestructura, plataforma o serveis analítics del model públics per millorar encara més l'escalabilitat; o implementar un model híbrid que combina els recursos i serveis en el núvol tant privats com públics.

## 6.3. Big Data Analytics com a Servei (BDaaS)

La *International Telecommunication Union* en la seva recomanació *ITU-T Y.3600* defineix *BDaaS* (*Big Data as a Service*)[47] com una categoria de servei en el núvol en la qual les capacitats que es posen a disposició del client del servei en el núvol li permeten recopilar, emmagatzemar, analitzar i visualitzar les dades utilitzant tecnologies Big Data.

Tot i que empreses que ofereixen aquests serveis també es refereixen a *Analytics-as-a-service* (*AaaS*) per descriure models basats en núvol per *Big Data Analytics*.

De moment, *BDaaS* és un terme un poc difús[48] sovint utilitzat per descriure una àmplia varietat d'externalització de diverses funcions de grans volums de dades al núvol. Aquest pot anar des del subministrament de dades, fins al subministrament d'eines analítiques amb les que interrogar les dades (sovint a través d'un panell d'instruments o de control en format web) per dur a terme l'anàlisi i l'elaboració d'informes. Alguns proveïdors de *BDaaS* també inclouen consultoria i serveis d'assessorament dins dels seus paquets *BDaaS*.

El seu model de negoci, igual que la de computació en núvol, es basa en el lloguer de serveis sota demanda de computació i / o emmagatzematge en mode pagament per ús. Cobra una important rellevància, en estalvi de costos en infraestructura *on-premise* i en personal qualificat, en les àrees on es processen dades massives mitjançant algoritmes de *Data Analytics*.

El *stack* o pila de serveis *BDaaS* està compost per nivells de grups de tecnologies d'acord a les funcions que exerceixen (Figura 11).

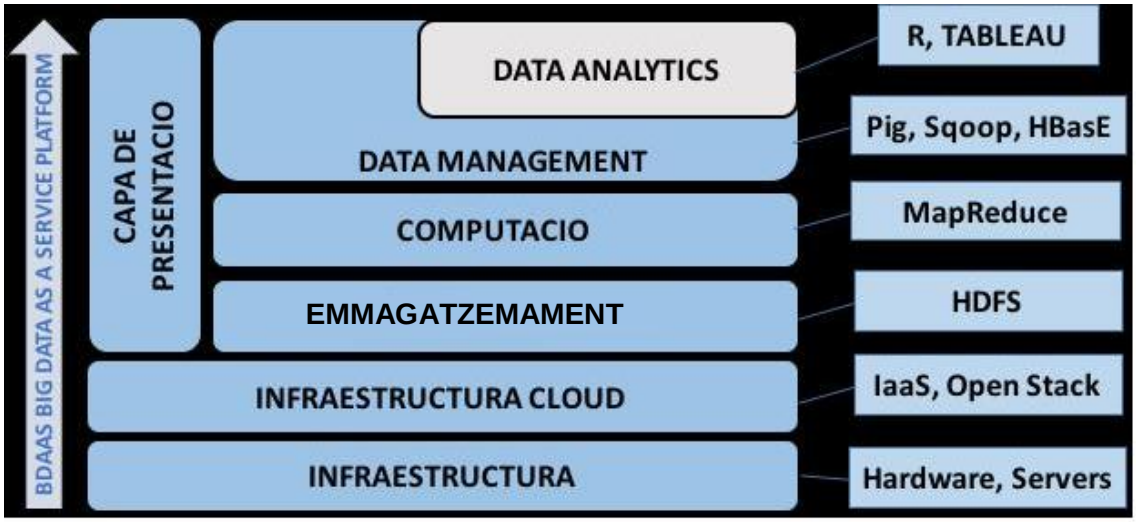


**Figura 11. Plataforma Big Data com a servei**[49]

- La capa *Data Analytics* inclou aplicacions analítiques d'alt nivell, com *R*, *Tableau* o *Tibco Spotfire*[50], sobre una solució *Cloud computing* que s'utilitza per analitzar les dades. L'interessant d'aquest model és que el tipus d'eina a utilitzar es pot especialitzar per a cada tipus d'indústria. Per exemple, per a entorn financer, minoristes o assegurances. La capacitat d'especialitzar la capa de *Data Analytics* dins la pila *BDaaS* fa que sigui utilitzable i adaptable a moltes organitzacions.

- La capa de *Data Management* inclou tecnologies de maneig de dades, com poden ser les pròpies de l'ecosistema *Hadoop Pig*, *HbasE*, *Sqoop*, etc.

- La capa de computació proveeix els serveis de còmput, i pot estar basada en *Amazon MapReduce*, o bé les més noves com *Spark*.

- La capa d'emmagatzematge disposa de la infraestructura necessària per a la implementació de *HDFS*, podent estar basada en *S3* d'*Amazon* o un altre model d'emmagatzematge distribuït i redundant.

- Finalment, en la capes de *Cloud* i infraestructura es pot optar per les solucions *IaaS* més comuns, com *VMWare*, *OpenStack* i estar ubicats a *Datacenters* distribuïts.

En aquest model de plataforma, els nivells més alts (Emmagatzemament, Computació, *Data Management* i *Data Analytics*) tenen una capa de presentació que possibilita als usuaris l'accés als serveis. Alguns exemples de companyies que ja es defineixen en aquest model són *Cazena*[51], *Qubole*[52] o *Doopex*[53].

## 6.4. Tipus de Servis en núvol per BDaaS

*BDaaS* es pot implementar en el núvol i es pot basar en diversos tipus de serveis. La determinació de la combinació adequada dels serveis depèn de les necessitats del client que els haurà de sospesar davant els recursos interns existents i el projecte concret de Big Data que pretén posar en marxa.

Els tipus de serveis en núvol bàsics per a l'analítica de dades massives com a servei inclouen la *Infraestructura com a Servei* (*IaaS*), *Plataforma com a Servei* (*PaaS*) i *Software com a Servei* (*SaaS*). Tot i que *IaaS* queda fora de la capa de representació per aquest nou model pot ser d'especial interès si es vol tenir el control total sobre les tecnologies Big Data a implementar.

### *6.4.1. Infraestructura com a Service (IaaS)*

*IaaS* proporciona la base per a serveis en el núvol de moltes empreses. No obstant això, *IaaS* també requereix una major inversió dels recursos de TI en el context de l'aplicació Big Data Analytics. El client serà responsable d'instal·lar el seu propi programari, com ara el marc de *Hadoop*, o una base de dades *NoSQL*, com *Apache Cassandra , MongoDB*, o tecnologies *Couchbase*. El client també serà responsable de la gestió dels seus recursos assignats mitjançant eines automatitzades que faciliten la gestió i l'orquestració de recursos.

Es podrà desplegar *on-premise* o a través d'un proveïdor del núvol, *IaaS* permet assignar o comprar temps en els recursos compartits de servidors, que sovint estan virtualitzats, per gestionar la capa de computació i les necessitats d'emmagatzematge per a l'anàlisi de grans volums de dades. Els sistemes operatius que proporciona *Cloud Computing* gestionen conjunts de servidors d'alt rendiment, la xarxa i els recursos d'emmagatzematge físic.

A continuació és mostren exemples de solucions *IaaS* dels proveïdors en l'ecosistema de la tecnologia de núvol:

- *Amazon* Web Services*
- *Citrix* CloudPlatform*
- *Windows Azure** i *Microsoft* System Center*
- *OpenStack* software*
- *Rackspace**
- *VMware vCloud* Suite*

### *6.4.2. Plataforma com a Servei (PaaS)*

*PaaS* proporciona als desenvolupadors les eines i biblioteques per a construir, provar, implementar i executar aplicacions en la infraestructura de núvol. D'aquesta manera redueix la càrrega de treball de gestió, eliminant la necessitat d'instal·lar i ampliar els elements de la seva implementació *Hadoop*, és a dir, la plataforma ja implementa *HDFS* i *Map/Reduce*, i serveix com a base per el desenvolupament d'aplicacions analítiques avançades.

El client d'aquest model no administra ni controla la infraestructura de núvol subjacent incloent la xarxa, servidors, sistemes operatius o l'emmagatzematge, però té control sobre les aplicacions que implementa i algunes configuracions d'entorn de la plataforma proporcionada. Per exemple, OpenShift[54] integra la plataforma de *Hortonworks* i la ofereix amb aquesta modalitat perquè els seus clients puguin desenvolupar aplicacions analítiques ràpidament.

La següent és una mostra de les solucions dels proveïdors de *PaaS* en l'ecosistema de la tecnologia de núvol.

- *Force.com*
- *Google* App Engine*
- *Red Hat* OpenShift**
- *VMware Cloud Foundry*
- *Windows Azure**

### *6.4.3. Software com a Servei (SaaS)*

Amb el model *SaaS*, la capacitat oferta al client és utilitzar les aplicacions del proveïdor que s'executen en una infraestructura de núvol. Les aplicacions són accessibles des de diversos dispositius client a través d'una interfície de client lleuger, com un navegador web. El consumidor no gestiona ni controla la infraestructura de núvol subjacent incloent la xarxa, servidors, sistemes operatius, emmagatzematge, plataforma, ni les capacitats d'aplicacions individuals, amb la possible excepció de valors de configuració que el proveïdor proporciona a les aplicacions específiques d'usuari.

Les aplicacions específiques per a l'analítica de Big Data basades en el núvol es poden aprovisionar amb *SaaS*. Podria ser necessari utilitzar múltiples aplicacions *SaaS* per cobrir la gamma d'escenaris que els usuaris de negocis requereixen. Per exemple, el programari que funciona bé per a l'anàlisi d'opinions pot no funcionar per a la gestió del risc o rendiment dels actius. *SaaS* pot ser ofert com una aplicació independent o com a part d'una solució més gran d'un proveïdor del núvol. Per exemple, *Jaspersoft for AWS*[55] utilitza el servei *Amazon Elastic MapReduce,* en seu servidor *cloud analytics server*, que ofereix en mode de pagament per ús, i que inclou eines avançades per la captura, anàlisi i visualització de dades.

Exemples de proveïdors en l'ecosistema de la tecnologia de núvol de les solucions SaaS de Big Data Analytics són:

.

- *Amazon* Elastic MapReduce*
- *Cetas* by VMWare* analytics solutions*
- *Google* BigQuery services*
- *Rackspace* Hadoop* service*
- *Windows Azure* HDInsight**

## 6.5. Avantatges de BDaaS

*Hadoop* de moment pareix la pedra angular per Big Data, però encara té les seves limitacions, especialment per a les petites i mitjanes empreses que no compten amb els recursos per construir una infraestructura *Hadoop on-premise*. A més, ho tenen més difícil per disposar de personal qualificat. *Big Data Analytics com a Servei* (*BDaaS*) s'ocupa d'aquestes qüestions, prenent l'avantatge de la flexibilitat de la computació en núvol. En aquest sentit, es poden enumerar les següents:

- Posada en marxa en qüestió de minuts. Iniciar un projecte de *Hadoop on-premise* pot prendre diversos mesos des del moment en què una empresa cerca opcions, l'implanta i el posa en marxa. Amb la computació en núvol, tot el que client ha de fer és seleccionar un proveïdor que proporcioni *Hadoop* amb alguna de les modalitats de servei i en pocs minuts es té accés a tota la infraestructura.

- Model de preus assequibles. El model de núvol permet a les empreses utilitzar espai d'emmagatzematge i potència de càlcul sota demanda a través d'un proveïdor extern, sent un model eficient i rendible. D'aquesta manera, no es requereix l'adquisició de nous recursos interns per a certes iniciatives d'anàlisi de grans volums de dades (com per exemple, els projectes a curt termini), on es pot proporcionar capacitat i escalabilitat addicional segons sigui necessari en mode de pagament per ús, i una vegada que el projecte estigui acabat deixar de pagar per aquest servei.

- Mantenir les dades sensibles internes. Amb una implementació de núvol híbrid, les empreses poden mantenir les seves dades més sensibles internament i incorporar enormes volums de dades (tant propietat de l'organització com generades per tercers o proveïdors públics) que no són tan sensibles al núvol públic per ser analitzades externament. Moure grans quantitats de dades a les instal·lacions pròpies pot ser un compromís significatiu de recursos.

- Augment del valor afegit. Big Data en el núvol pot eliminar la necessitat de trobar un científic de dades o expert en TI per ajudar a implementar un projecte de Big Data. En molts de casos, el proveïdor ofereix eines que són fàcils d'usar i proporciona suport tècnic per a qualsevol problema que pugui trobar-se. Els departaments TI en lloc de dedicar-se en les tecnologies d'infraestructura de Big Data, poden centrar-se en aspectes de més alt nivell, com és obtenir informació útil per el seu model de negoci.

Mitjançant *BDaaS*, segurament seran les petites i mitjanes empreses i nínxols departamentals o funcionals, els que es beneficiaran més d'aquest model de presentació Big Data ja que podran disposar d'aquesta manera de capacitats analítiques sense haver d'incórrer en l'exorbitant despesa que comporta tenir empleats a temps complet o l'ampliació que podria suposar la implementació de *Hadoop* en un centre de dades *on-premise*.

Tampoc tots els sectors es beneficiaran per igual del Big Data. Tot i que és molt probable que totes aquestes dades disponibles per ser explotades cada vegada adquireixin més importància en el dia a dia dels negocis. Els sectors que probablement es beneficiaran més del *BDaaS* són: les institucions financeres, el comerç minorista, l'administració pública i la investigació mèdica.

Concretament, en la investigació mèdica, a part d'estalvis d'infraestructura, també possibilita l'agrupament de dades a escala global. Això permetria anàlisis estadístics més precisos que millorin la comprensió i el coneixement d'aspectes de la salut i que repercuteixen en millors diagnòstics i tractaments.

# 7. Aplicacions BDaaS en Sanitat

L'anàlisi del Big Data ha obert la porta a una nova era per a la millora en la prestació de serveis i solució de problemes en l'àmbit dels sistemes sanitaris. La gran majoria dels agents que participen en les estructures dels serveis de salut reconeixen que l'anàlisi del Big Data pot oferir noves possibilitats en l'elaboració de models predictius, patrons de comportament, el descobriment de noves necessitats, reduir riscos, així com proveir serveis més personalitzats, tot això en temps real i tenint en compte tota la informació rellevant.

Segons la infografia de Siemens, mostrada en el capítol 5 i que també defineix la publicació de *Feldman. B. "Big Data in Healtcare Hype and Hope" [38],* agrupa en sis grups a empreses i organitzacions que estan començant a aprofitar Big Data per fer front a diferents desafiaments en el sector de la salut, el que reflecteix sis formes on Big Data pot ajudar a l'assistència sanitària.

Aquestes són: suport a la recerca, transformació de dades en informació, suport a l'autocura, suport als proveïdors de salut per millorar l'atenció als pacients, augmentar el coneixement, posada en comú de les dades per un ecosistema millor. Es descriuen amb més detall en els següents apartats posant com exemple aplicacions de proveïdors que ofereixen els seus serveis analítics a través del núvol, seguint el context i justificació del treball.

## 7.1. Suport a la Recerca

La Genòmica ha estat l'avantguarda de la revolució de grans volums de dades en les ciències de la salut, ja que sosté la promesa considerable de possibilitar la medicina personalitzada. A part de la que es comenta hi ha més empreses que es centren en la genòmica, cada un d'elles està prenent un enfocament diferent a les dades, amb l'esperança d'accelerar la investigació i en última instància, transformar el desenvolupament del tractament i la pràctica mèdica.

### *7.1.1. 23andMe*

Finançada -entre altres empreses-, per *Google*, *23andMe*[56] ofereix des del seu naixement, en 2006, test genòmics a l'abast de qualsevol. *23andMe* realitza proves genètiques a través d'una mostra de saliva i assegura que un 80% dels clients accepta que les seves dades s'utilitzin de forma anònima en projectes d'investigació.

L'empresa, que té una base immensa de dades amb l'ADN dels seus clients, utilitza la informació anònima del voltant d'1 milió persones també per al desenvolupament de nous fàrmacs gràcies a acords amb diverses empreses farmacèutiques. Tot i que s'ha topat amb el recel de l'Administració d'Aliments i Medicaments dels Estats Units (FDA), que demana que s'informi de manera més clara als seus clients, i els seus test estan prohibits en diversos països europeus. *23andMe* segueix creixent i el seu últim gran acord ha estat la venda

de 14.000 perfils genètics de malalts de Parkison o familiars a l'empresa *Genentech* per 60 milions de dòlars.

## 7.2. Transformació de dades a la Informació

Donada la creixent allau de dades de salut, la gran necessitat no satisfeta és gestionar millor aquestes dades. Un aspecte clau és la transformació de dades en informació útil. Estructurar les dades no estructurades per a la gestió computacional és una important trampolí per permetre l'assistència sanitària basada en dades. Tal vegada intuïtivament, en alguns casos, convertir la informació no estructurada (històries clíniques i les notes mèdiques) en dades és un primer pas necessari. L'analítica explicativa utilitza un conjunt d'eines basades en la mineria de dades, anàlisi de conglomerats, estadístiques, visualitzacions de dades, màquines d'intel·ligència artificial, anàlisi de texts, i el processament del llenguatge natural (NLP) per extreure patrons i significat.

### *7.2.1. Proscia*

*Proscia*[57] és una companyia nord-americana que ha creat una gran base de dades a nivell mundial on s'analitzen imatges i dades de tot tipus de patologies, però el seu principal objectiu és la lluita contra el càncer. Es defineix com la "primera plataforma en el núvol per la patologia digital". Va er desenvolupada inicialment per diferents universitats americanes.

*L'oferta de Proscia* és ajudar als professionals de la salut a diagnosticar, tractar i prevenir el càncer, combinant científics de dades en el núvol, analitzant imatges i utilitzant *machie learning* amb algoritmes intel·ligents. Les característiques tècniques del sistema són, entre altres, un emmagatzematge segur, anotació i col·laboració en biòpsies digitals amb alta resolució, visualització i profunditat del zoom.

La plataforma *Proscia Pathology Cloud* accelera els seus processos d'emmagatzematge avançat, col·laboració i capacitats d'anàlisi. A més aprofita el potencial del núvol amb la còpia de seguretat de les dades escanejades, protecció de dades, i recull les percepcions dels seus socis en temps real. La plataforma pretén organitzar tota la informació patològica a nivell mundial, per això, registrar-se no té cap cost i l'ús del programari és gratuït fins als 20 GB d'emmagatzematge.

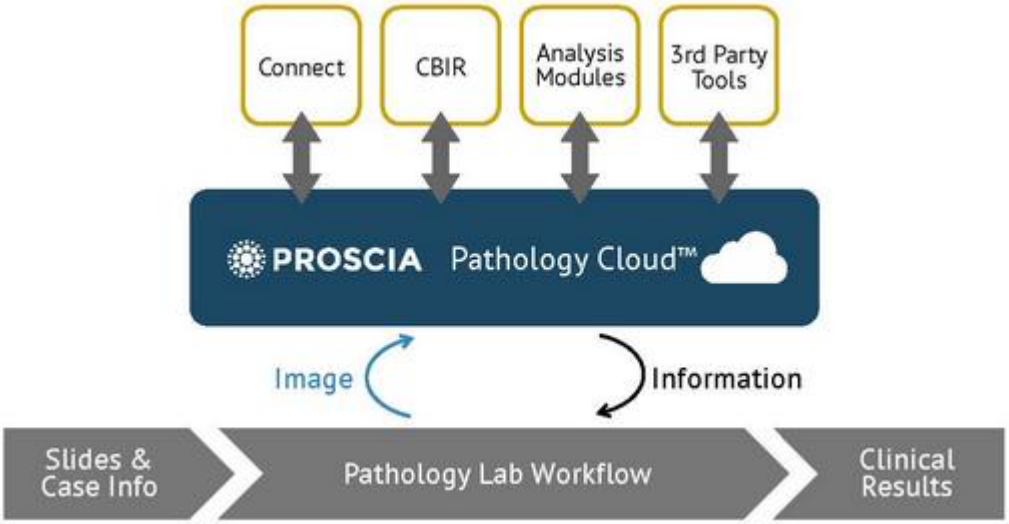


**Figura 12. Pathology Cloud Platform**[58]

La plataforma per la patologia digital resol varis obstacles que fan difícil per a moltes institucions d'aquest sector el disposar d'eines analítiques avançades, a més de l'escalabilitat, seguretat i rendibilitat. El programari com a servei de *Proscia*, recolzat per *Amazon Web Services* (AWS), proporciona als clients una forma única d'emmagatzematge segur i escalable i solució d'allotjament que pot satisfer les necessitats de qualsevol institució.

### *7.2.2. PatientsLikeMe*

*PatientsLikeMe*[59] és una de les xarxes socials de salut més conegudes en l'actualitat. Fundada per *Jamie Heywood* que, després de veure com al seu germà se li diagnosticava ELA (esclerosi lateral amiotròfica), va posar en marxa primera plataforma 2.0 on milers de pacients comparteixen les seves dades clíniques.

*Heywood* assenyala que desenes de milers de vides es perden cada any a causa de que les dades de salut no flueixen lliurement i ha aconseguit que en *PatientsLikeMe* comparteixin informació més de 125.000 persones afectades per malalties greus i incurables com la SIDA, càncer, diabetis o ELA. Actualment estan representades més de 1.000 malalties.

Els pacients poden, des de la web, seleccionar un dels grups d'interès de la xarxa social i introduir les seves dades, comptar l'evolució dels símptomes, quins efectes tenen els tractaments (fins i tot els que prenen de manera irregular, sense que se'ls recepti un metge), els trucs i l'avanç de la malaltia, indicant com influeixen els medicaments en el seu benestar físic, mental i social; les hospitalitzacions que pateixen, el seu pes o estat d'ànim. A partir d'aquí, es comença a monitoritzar l'evolució de la seva malaltia comparant-se amb altres pacients. El web també ofereix als seus usuaris, de manera gratuïta, eines analítiques de gran potència que abans només estaven a l'abast dels investigadors.

A diferència d'altres xarxes socials, a '*PatientsLikeMe*' no existeix la privacitat

de dades. Qualsevol usuari pot accedir a informació mèdica d'un altre usuari, així com altres aspectes de la seva vida com l'historial familiar, el seu estat d'ànim o les seves conviccions religioses. És el que a '*PatientsLikeMe*' en diuen *'Open Philosophy'*. De fet, el model de negoci d'aquesta xarxa social de salut sorgeix a partir d'aquesta filosofia. Els responsables de la plataforma venen dades agregades, anònims i segmentats a les companyies farmacèutiques. I al mateix temps ajuden a millorar la qualitat de vida dels pacients i donen suport a la investigació mèdica.

A més, gràcies al sistema *'Open Philosophy'* es pot obtenir un tipus d'informació sobre pacients que, fins al moment, era pràcticament impossible d'aconseguir, fins i tot per institucions dedicades a la investigació. A part d'això, l'elevada mostra poblacional permet estudis d'investigació mèdica molt fiables a nivell estadístic.

## 7.3. Suport a l'autocura

Un altre grup d'empreses està utilitzant Big Data en les noves formes d'ajudar a als pacients a tenir cura d'ells mateixos. Combinant la comoditat dels telèfons mòbils amb el poder de grans volums de dades, permeten recopilar informació mèdica i el seguiment de la son, mentre que els malalts crònics poden gestionar millor la diabetis, malalties del cor i asma. L'anàlisi d'aquestes dades permet entendre millor els patrons de conducta i motivacions per realitzar canvis que permetin prevenir, impedir o mitigar malalties.

### *7.3.1. SmartWatch per la salut*

*Fitbit*[60] es un dels monitors d'activitat física més populars. Orientat al consum de masses es comercialitzen diferents dispositius adaptats segons els seus objectius. El més complet registre localitzacions, freqüències cardíaques, el temps que es dorm i la seva qualitat, calories cremades, passes i trams d'escales pujats, distancia i temps d'activitat. Apple Wach[61] també implementa funcions similars.

Aquests dispositius poden comunicar les dades amb aplicacions de salut compatibles que gestionarà les dades per la seva visualització i opcionalment la seva compartició.


**Figura 13. Monitors Fitbit**

## *7.3.2. ECG SmartWatch*

*Smartphone* amb ECG de 6 derivacions i ECG *SmartWatch*. Encara que una sola derivació està disponible des de fa uns anys, es tracta d'un important avanç en la recerca d'aconseguir un millor diagnòstic en afeccions cardíaques mitjançant l'ús de múltiples derivacions o, en el cas del rellotge, el reconeixement immediat i la notificació a partir de canell de l'usuari.

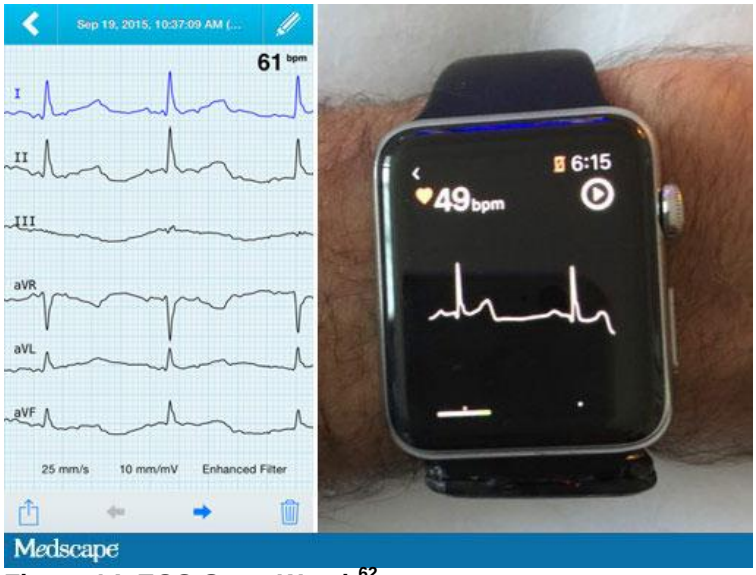


**Figura 14. ECG SmartWatch**[62]

## *7.3.3. Nightwatch*

Permet mesurar els nivells de glucosa amb un *Smartphone* o *SmartWatch*. Són molt útils per al seguiment de forma contínua dels nivells de glucosa, amb només mirar a la pantalla de telèfon o rellotge, un exemple es *Nightwatch*[63], un 'app' per *Android Wear* que és capaç de connectar-se a un monitor continu de glucosa (concretament, a un *Dexcom CGM*) i rebre informes actualitzats cada

cinc minuts, proporcionant tendències en intervals de 3, 6 o 24 hores i amb alertes de nivells preestablerts.

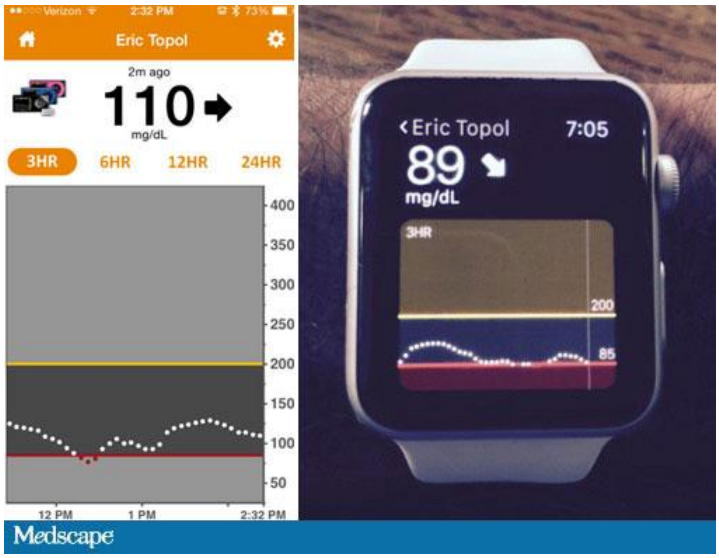


**Figura 15. Monitor de glucosa [63]**

## 7.4. Suport a Proveïdors, Millora de l'Atenció al Pacient

Els proveïdors de salut s'enfronten a un augment de pressió: menys temps i diners per fer més amb un caudal cada vegada més gran d'informació, i tot això, sense deixar de proporcionar una bona atenció als pacients i no cometre errors. El suport al proveïdor és una de les àrees més interessants en que Big Data pot ajudar. No obstant això, la resistència al canvi dels proveïdors i interfícies d'usuari poc amigables segueixen sent els principals reptes en aquest camp. Diverses companyies estan prenent diferents enfocaments per a la construcció de sistemes de suport al proveïdor que són fàcils d'usar, permeten estalviar diners i millorar els resultats alhora que els proveïdors disposen de més temps per a donar una bona atenció als pacients.

### *7.4.1. Ecògraf en un Smartphone*

*Lumify*[64] de *Philips* permet incorporar una sonda portable en un telèfon *Android* i obtenir imatges d'ultrasò d'alta qualitat, que poden enviar-se fàcilment al pacient o a un especialista per a la interpretació ($ 199 / mes - il·limitat d'usuaris).


**Figura 16. Lumify de Philips**

### *7.4.2. PillCam*


Les proves mèdiques no es queden fora d'aquesta revolució. La PillCam[65], ja aprovada per al seu ús en EE.UU, es compon d'una bateria, una llum i dues minicàmeres que prenen 4 imatges per segon (2 per cada càmera) i pot capturar imatges durant 9 o 10 hores. Es usada com a alternativa a la colonoscòpia per a pacients que no hagin pogut completar-la. Transmet les imatges a una gravadora que el pacient porta al cinturó.

**Figura 17. PillCam**

## 7.5. L'augment del coneixement

Big Data és un conjunt d'eines òbvies per augmentar el coneixement per començar a resoldre una varietat de problemes basades en dades: identificació de la falsificació de medicaments, el seguiment dels problemes ambientals que desencadenen l'asma, la predicció de brots de malalties, ajudant als països en desenvolupament a prendre millors decisions polítiques; tot això amb la finalitat de donar prioritat a les qüestions de salut pública mundial.

### *7.5.1. Medisys*

La Comissió Europea ha desenvolupat el sistema *MediSys*[66], una eina per escanejar i buscar informació per reformar la xarxa de vigilància de malalties infeccioses i la seva detecció primerenca. A través de *MediSys* es poden obtenir notícies utilitzant més de 20.000 articles d'Internet que s'analitzen cada dia i són produïts per l'*Europa Media Monitor*, i que es poden enviar a les persones clau, gestors, decisors, etc., per correu electrònic i *SMS*.

### *7.5.2. Propeller Health*

*Propeller Healt* es un inhalador que porta un *GPS* incorporat on els sensors de localització mesuren el temps i el lloc de cada cop que es fa ús del inhalador. Aquest dispositiu –que encaixa en qualsevol inhalador ordinari- està dissenyat per ajudar els pacients amb malalties respiratòries cròniques i als seus metges a controlar la malaltia (asma, *MPOC*, altres). Permet realitzar un seguiment dels símptomes, els factors desencadenants, l'ús de medicaments i compartir la informació amb els metges.

*Propeller Health*[67] ha llençant el seu projecte en *Louisville* (EEUU) on centenars de pacients poden ajudar a localitzar “punts calents” en els que és difícil respirar. Basant-se en les dades obtingudes, la companyia espera fer recomanacions als planificadors urbans per mitigar els problemes.


**Figura 18. Propeller Healt**

## 7.6. Posada en comú de dades per un Ecosistema de millor

Una interessant aplicació de Big Data és ajuntar conjunts de dades dispars en formes que puguin permetre nous tipus d'anàlisi i facilitar les respostes a les grans preguntes, algunes de les quals encara no s’han plantejat.

### *7.6.1. IBM Watson Healt Cloud*

*IBM Watson Health* ofereix una plataforma al núvol[68] oberta i segura per a metges, investigadors, companyies asseguradores i companyies orientades a solucions en salut i benestar (*IBM Watson Health Cloud*) que permet anonimitzar, compartir i combinar les dades referents a la salut. La seva utilització permetrà als professionals comptar amb una percepció més completa dels múltiples factors que poden afectar la salut del pacient.

En aquesta iniciativa, *IBM* compta amb la col·laboració de companyies líders en els seus sectors com *Apple, Johnson & Johnson* i *Medtronic* que optimitzaran els dispositius mèdics i de consum per a la recopilació de dades i la seva posterior anàlisi. A més, *IBM* ha adquirit les companyies de tecnologia sanitària *EXPLORIS* i *Phytel*, especialitzades en analítica.

### *7.6.2. IBM Watson Care Manager*

En aquest sentit *IBM* ha desenvolupant *IBM Watson Care Manager*[69], una nova solució per a l'àmbit de la salut que incorpora la tecnologia de: Phytel, programari que facilita la gestió i anàlisi de dades personalitzades sobre la salut des del núvol; i l'*Apple Health Kit* i el *Research Kit* d'*Apple*, programari dissenyat per *Apple* per facilitar que els investigadors puguin desenvolupar els seus estudis utilitzant els seus *iPhone*. Aquesta nova solució integra diferents tipus de dades clíniques individuals, als quals s'aplica anàlisi cognitiva per extreure conclusions que puguin ajudar els metges i infermers a fer un seguiment detallat dels pacients on les condicions mèdiques són complexes i sovint costoses.

Per exemple, un pacient amb una malaltia crònica de cor que obté l'alta hospitalària requereix un seguiment personalitzat des de casa, que inclou la comprovació diària del seu pes i la seva activitat física. Tant l'enviament de dades del pacient al metge com l'anàlisi que aquest realitza s'ha fet des de fa anys de forma manual. Gràcies a *IBM Watson Care Manager* un pacient podrà rebre en el seu dispositiu -per exemple, un *Apple Watch*- dades recollides per diferents tipus de sensors, via *wireless* o dispositius "*wearables*". Els professionals de la salut podran rebre conclusions derivades de l'anàlisi cognitiva del flux de dades d'un pacient individual, amb l'objectiu de prevenir problemes mèdics. Les dades individuals dels pacients es podran bolcar a *IBM Watson Health Cloud* on s'aniran analitzant a mesura que avança el temps per extreure coneixement que pugui millorar els tractaments.

Recentment, aquesta divisió d'*IBM* disposa d'un equip de 2.000 professionals que tindrà la responsabilitat d'expandir el negoci a escala global, així com ampliar les capacitats d'*IBM Watson Health* en el núvol i fer créixer l'ecosistema al voltant d'aquesta plataforma tecnològica.

### *7.6.3. IBM Care Management*

Cal destacar, d'entre els productes mes madurs d'IBM per sanitat, ***IBM Care Management***[70] que és una aplicació de programari empaquetada que ofereix les capacitats clau necessàries per gestionar l'assistència en l'atenció continuada. Permet identificar clients que necessiten assistència, avaluar les seves necessitats, establir el pla d'assistència adequada per respondre a les seves necessitats, a més de gestionar l'assistència i supervisar els resultats obtinguts.

*IBM Care Management* combina la integració de dades, les analítiques i la coordinació de les capacitats d'assistència en una única oferta llesta per utilitzar que aporta una visió completa i personalitzada de l'individu a fi de facilitar una assistència centrada en els resultats. En la seva oferta, agrupa les funcionalitats en tres grups.

1. Enfocament centrat en el pacient i basat en l'equip per a l'assistència a través de la col·laboració de tota l'organització:

- Coordinar i col·laborar en tot un equip multidisciplinari que pot incloure metges, infermeres, treballadors socials, assessors en salut mental, fisioterapeutes, treballadors comunitaris i membres de famílies.
- Visualitzar el perfil de client biopsicosocial en una visió completa.
- Rebre expedients utilitzant un flux de treball configurable i crear automàticament un pla de resultats.
- Utilitzar una interfície de planificació de resultats flexible i intuïtiva per crear plans d'assistència més complets que vagin més enllà dels tradicionals plans d'assistència clínica.
- Col·laborar entre les diverses parts interessada de forma eficient per coordinar l'assistència, localitzar i derivar als proveïdors d'assistència i optimitzar els recursos.

2. Aporta una vista única del pacient i del pla d'assistència utilitzant el suport d'integració basat en estàndards:

   - Suport a la integració bidireccional amb els EMR (registre mèdic electrònic) i altres sistemes d'origen de dades, seguint els estàndards per a l'intercanvi de dades.
   - Ús d'un conjunt d'eines de correlació gràfica, connectors (nodes) i IHE, HL7 i models de desenvolupament i esquemes continus per a una integració més fàcil.
   - Unifica i sincronitza la informació de salut mental, social i clínica per crear una vista única i personalitzada del pacient i del pla d'assistència.

3. Enriquiment de les dades de la solució de gestió de l'assistència identificant les dades de l'assistència mèdica en dades no estructurades i convertir-los en codis estàndards:

   - Identifica amb precisió les dades referides a l'assistència de les dades no estructurades introduïdes en la solució, com les notes del metge i les notes del treballador social.
   - Converteix les dades no estructurades en codis estàndard utilitzant acceleradors llestos per al seu ús.
   - Analitza la informació no estructurada que suma el 80 per cent de dades de salut i assistència social convertint-la en dades estructurades per obtenir un millor coneixement de l'individu.
   - Analitza les dades no estructurades per aportar informació de valor en l'àrea d'assistència on més es necessiti.
   - Aprofita més de 100 diccionaris i 800 regles d'anàlisi per extreure i correlacionar informació en els diagnòstics, procediments, laboratoris i medicaments utilitzant estàndards del sector com, per exemple, ICD-10, CPT, SNOMED i RXNORM

# Conclusions

Actualment són moltes les promeses i poques les aportacions de Big Data en el sector de la salut. Principalment són algunes grans empreses del sector de les TI i algunes *startups* que es centren nínxols funcionals les que estan impulsant l'evolució en els centres sanitaris. També comencen a aparèixer algunes organitzacions governamentals de l'àrea de sanitat que treuen profit de l'anàlisi de Big Data.

Big data és molt més que la mera descripció de les seves *Vs ,* requereix de grups de servidors orquestrats per suportar les eines que processen grans volums, a alta velocitat, i variats formats de dades. A part, per poder treure valor requereix d'amplis coneixements en enginyeria de dades, mètodes científics, matemàtiques, estadística, computació avançada, visualització i coneixements en els diferents àmbits d'especialitat on s'aplica. Calen així de professionals experts en aquests sector i que a més, és difícil de trobar en un sol individu.

Big Data Analytics no és el substitut de les solucions tradicionals d'intel·ligència de negoci. Cada sistema té les seves avantatges i febleses. El valor que pugui aportar un o l'altre depèn de la situació concreta en que s'ha aplicar. El volum, la varietat i la velocitat de les dades marquen la diferencia i inclús la seva combinació pot ser una solució apta.

Les tecnologies Big Data estan en continua evolució. Apareixen renovats sistemes de processament de dades que amplien les funcionalitats de les plataformes permetent realitzar anàlisis de dades en temps real o quasi real. Així mateix, una arquitectura de plataforma Big Data Analytics no és un model únic, pot implementar diferents components depenent del propòsit de la solució concreta a implementar.

El sector sanitari genera moltes dades. El 80 per cent d'aquestes dades són en format no estructurat que corresponen, entre altres, als registres mèdics electrònics, i la major part, a la imatge per el diagnòstic. Per analitzar aquestes dades de manera eficient és requereix de les tecnologies Big Data i de coneixements específics del camp sanitari així com els englobats en *Data Science*.

Les tecnologies Cloud Computing també segueixen evolucionant. Els seus models poden ajudar a accelerar el potencial de les solucions d'anàlisi escalables. El Cloud ofereix flexibilitat i eficiència per accedir a dades, el lliurament de coneixements, i l'obtenció de valor. No obstant això, Big Data Analytics en el núvol tampoc és una solució de talla única.

Les organitzacions que utilitzin la infraestructura de núvol per proporcionar Analítica *Big Data com a Servei* (BDaaS) tenen múltiples opcions. Per factors de càrrega de treball, el cost, la seguretat i la interoperabilitat de les dades, els clients poden optar per utilitzar el seu núvol privat per mitigar el risc i mantenir el control; o utilitzar infraestructura, plataforma o serveis públics d'anàlisi per

millorar encara més l'escalabilitat; o implementar un model híbrid que combina els recursos i serveis en els núvols privat i públic.

Els avantatges que pot aportar l'anàlisi Big Data a la sanitat no deixen indiferent, tant pel que fa a la reducció de costos però, principalment, en els beneficis que promet per la nostra salut. Tot i així, perquè sigui una realitat s'hauran de superar les barreres administratives, tecnològiques i, el gran taló d'Aquil·les, la seguretat i privacitat de les dades personals.

Pot ser el futur, per l'obtenció de valor rellevant a partir de l'analítica de les dades sanitàries, està en combinar les tecnologies de Big Data Analytics i Cloud Coputing. Mitjançant l'agrupament de dades anonimitzades a escala global en nínxols específics altament especialitzats per el seu anàlisi, i aquests correctament integrats amb els sistemes d'informació sanitaris, permetrien aconseguir un ecosistema que aportaria els beneficis esperats.

En referència als objectius plantejats inicialment, no s'han assolit completament. En el començament del projecte, durant la fase de recerca de fonts primàries i secundaries, tant per la quantitat de proveïdors que ofereixen sistemes d'analítica Big Data com per de les solucions sectorials per sanitat (moltes vegades poc especifiques) i l'enorme publicitat que envolta aquest concepte varen fer agafar un caire optimista que va fer menysprear l'amplitud i profunditat del concepte Big Data, la complexitat de la seva arquitectura i la problemàtica que comporta analitzar dades no estructurades. Això va fer endarrerir una part del treball que s'havia d'incloure en la PAC2.

Així mateix, l'objectiu en la PAC3 era definir un projecte Big Data per sanitat, però quant es va entrar en més en detall en aquest sector, es va considerar que definir un projecte concret no era una bona opció principalment per dos motius. Per una part, es requerien més coneixements en certs aspectes de les '*Ciències de les Dades*' i de l'àmbit sanitari fins a un cert nivell de profunditat. Per altra part, la impossibilitat d'abastar tot aquest coneixement i complir amb les tasques i dates de lliurament.

La planificació corresponent a la PAC2 es va seguir en els passos inicials, en l'estudi de Tecnologies Big Data i en l'anàlisi dels seus components i capacitats. En aquest punt, el contingut que s'analitzava va resultar ser més elevat del previst. D'aquesta manera, no es podia fer una comparativa de proveïdors de Big Data ja que no es tenien el conceptes de la seva arquitectura prou clars.

Pel que fa a la planificació de la PAC3 es canvia l'objectiu principal de definir un projecte Big Data per sanitat a favor d'un enfocament més generalista que inclou l'anàlisi necessari del sector sanitari per presentar una arquitectura Big Data genèrica que pot abastar la totalitat de les dades sanitàries i exemples d'aplicacions que, en aquest sentit, encaixen en el context i justificació del treball. Així mateix, dins el context del treball, també es realitza una anàlisi de les plataformes Big Data sobre tecnologies Cloud Computing, Tot i aquests canvis, es segueixen les tasques planificades en quant a l'agregació d'aspectes

mancants de la PAC2, anàlisi de riscos, barreres i protecció de dades personals en el sector sanitari.

D'aquesta manera, la planificació inicial no era prou adequada. La diferencia entre els coneixements que es tenien i els que es requerien per la consecució dels objectius inicials no era realista per la impossibilitat d'abastar tot el coneixement amb el temps marcat. Així els canvis introduïts en la planificació han possibilitat una visió més general de totes les àrees a tractar. Tal vegada, ara, amb els coneixements adquirits, es podria pensar en afrontar les fases de planificació i construcció d'un escenari per un projecte Big Data per sanitat concret.

# Glossari

**BI:** igual que Business Ingelligence.

**Big Data**: dades caracteritzades per el volum, varietat i velocitat.

**Big Data Analytics**: sistema per aplicar intel·ligència de negoci a les dades massives.

**Business Intelligence**: sistema que analitza les dades de l'empresa i mostra com està funcionant el negoci en les seves diferents àrees per poder prendre les millors decisions.

**CIM / CIE / ICD:** és la "Classificació Internacional de Malalties", en castellà CIE i en angles ICD ( "*International Statistical Classification of Diseases and Related Health Problems*") i té com a objectiu classificar i codificar malalties, signes, símptomes, troballes anormals, denúncies, circumstàncies socials i causes externes de dany o malaltia per recopilar informació sanitària útil relacionada amb defuncions, malalties i traumatismes (mortalitat i morbiditat). Ha passat per diverses versions. L'última d'elles és ICD-10, publicada el 1992, encara que la versió 9 segueix sent utilitzada.

**Cloud Computing**: (computació al núvol) És un paradigma que permet oferir serveis de computació a través d'una xarxa, que usualment és Internet. Els seus models principals de desplegament i de servei estan descrits en el capítol 6.

**CPT**: ("*Current Procedural Terminology*") és un catàleg codis mantingut per l'Associació Mèdica Americana. Aquest catàleg descriu els serveis mèdics, quirúrgics i de diagnòstic per tal d'unificar aquesta informació entre metges, codificadors, pacients, institucions, organitzacions d'acreditació i administradors. Aquesta codificació unificada permet homogeneïtzar tasques administratives, financeres i analítiques en el sector.

**Dades massives**: igual que Big Data.

**DICOM**: ("*Digital Imaging and Communication in Medicine*") és un estàndard per al maneig, emmagatzematge, impressió i transmissió d'imatges mèdiques que inclou la definició d'un format de fitxer i d'un protocol de comunicació en xarxa basat en TCP/IP.

**EMR**: (registre mèdic electrònic), registre amb informació clínica del pacient. En utilitzar aquest terme es sol referir més generalment a informació mèdica de pacients que no s'integra completament, o no s'integra, amb la HCE.

**MPOC**: (malaltia pulmonar obstructiva crònica)

**HCE**: (història clínica electrònica), també anomenada història clínica informatitzada (HCI), és el registre informatitzat de les dades socials,

preventives i mèdiques d'un pacient, obtingudes de manera directa o indirecta i actualitzades.

**HL7**: "*Health Level Seven*" és un conjunt d'estàndards per a l'intercanvi electrònic d'informació clínica, s'utilitza per als dominis clínic, assistencial, administratiu i logístic per aconseguir una interoperabilitat real entre diferents sistemes d'informació en el àrea de salut.

**IHE**: ("*Integrating the Healthcare Enterprise)*" és un conjunt d'especificacions que formen un marc tècnic com a recomanació d'ús d'estàndards existents. IHE és una iniciativa de professionals de sanitat i empreses per millorar la comunicació entre els sistemes d'informació que s'utilitzen en l'atenció al pacient. Defineix perfils d'integració per aconseguir la integració de sistemes amb interoperabilitat efectiva i flux de treball eficient a través d'estàndards que ja existeixen com HL7.

**Intel·ligencia de negoci**: igual que Business Intelligence.

**On premise**: en instal·lacions pròpies.

**RxNorm:** nom normalitzats de medicaments.

**SaaS**: (*Software as a Service), e*n català 'Programari com a Servei'. És un model de distribució de programari on el suport lògic i les dades que maneja s'allotgen en servidors d'una companyia de tecnologies d'informació i comunicació (TIC), als quals s'accedeix via Internet des d'un client. L'empresa proveïdora TIC s'ocupa del servei de manteniment, de l'operació diària i del suport del programari utilitzat pel client. Regularment el programari pot ser consultat en qualsevol computador, es trobi present en l'empresa o no.

**SNOMED-CT**: (*Systematized Nomenclature of Medicine - Clinical Terms*) és una terminologia clínica integral, multilingüe i codificada amb gran acceptació a nivell mundial. Permet fusionar termes en els àmbits de les ciències bàsiques, la bioquímica, les especialitats mèdiques i els continguts d'atenció primària creant una terminologia de referència que permet a nivell mundial la representació de la informació clínica de manera precisa i inequívoca.

# Bibliografia


[1] Fundación Rock Health. *"Big Data in digital Health";* octubre 2012*.* Accessible a: http://www.slideshare.net/RockHealth/rock-report-big-data, [consultat març, 2016]

[2] Parenteau J, Sallman R, Howson C, Tapadinhas J, Schlegel K, Thomas W. "*Magic Quadrant for Business Intelligence and Analytics Platforms"*: Gartner; febrer 2016. Accessible a: https://www.gartner.com/doc/reprints?id=1-2XXET8P&ct=160204, [consultat març, 2016]

[3] EMC Newsroom. "*Digital Universe Invaded By Sensors":* EMC; Abril 2014. Accessible a: http://www.emc.com/about/news/press/2014/20140409-01.htm [consultat març, 2016]

[4] Ericson Mobile Report, *"On The Pulse Of The Networked Society"*: Ericson; Agost 2014. Accessible a: https://www.ericsson.com/res/docs/2014/ericsson-mobility-report-august-2014-interim.pdf [consultat març, 2016]

[5] Barranco, R., *"¿Qué es Big Data?"*: IBM Software Group; juny 2012. Accessible a: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/, [consulta març, 2016]

[6] Visual Networking Index (VNI). *"VNI Mobile Forecast"*: Cisco; febrer 2016. Accessible a: http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html [consultat març, 2016]

[7] Gartner Newsroom. *"Gartnet Says 4.9 Billion Connected 'Things' Will Be in Use in 2015"*: Gartner; Novembre 2014. Accessible a: http://www.gartner.com/newsroom/id/2905717 [consultat març, 2016)

[8] Marr B. "*Big Data: 33 Brilliant And Free Data Sources For 2016*": Forbes; febrer 2016. Accessible a: http://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#2a1af6596796 [consultat març, 2016]

[9] Manyika J, Chui M, Brown B, Bughin J, Dobbs R., Roxburgh C, Byers A, *"Big data: The next frontier for innovation, competition, and productivity"* (disponible en pdf): McKinsey Global Institute, USA. Accessible a: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation, [consultat març, 2016]

[10] Gartner News Room: *"Gartner Survey Reveals That 73 Percent of Organizations Have Invested or Plan to Invest in Big Data in the Next Two Years"*: Gartner; setembre 2014. Disponible a: http://www.gartner.com/newsroom/id/2848718 [consultat març, 2016]

[11] Groves T. "*Where Does Hadoop Fit in a Business Intelligence Data Strategy?*": IBM; gener 2013. Accessible a: http://www.ibmbigdatahub.com/blog/where-does-hadoop-fit-business-intelligence-data-strategy [consultat abril, 2016]

[12] Mayer-Schönberger V, Cukier K. Big Data. "*La revolución de los datos masivos*". Turner Publicaciones S.L., Houghton Miffilin Harcourt, Madrid, 2013.

[13] Laney D. "*3D Data Management: Controlling Data Volume, Velocity, and Variety"*: Application Delivery Strategies, Meta Group; febrer 2001. Accessible a: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [consultat març, 2016]

[14] IBM Institute for Business Value. "*Analytics: el suso de big data en el mundo real*": IBM Global Business Services, IBM, 2012. Accessible a: http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf [consultat març, 2016]

[15] UPC School. "*Los profesionales más demandados: los expertos en Big Data*": Blog UPC School, UPC; novembre 2014. Accessible a: http://www.talent.upc.edu/blog/los-profesionales-mas-demandados-los-expertos-en-big-data/ [consultat abrir, 2016]

[16] Vazquez M. "*El científico de datos: un perfil multidisciplinario altamente especializado*": COMeIN, número 46; juliol 2015. Accessible a: http://www.uoc.edu/divulgacio/comein/es/numero46/articles/Article-Merce-Vazquez.html [consultat abril, 2016]

[17] Clasificaciones y normalización estadística. "*Transición a la CIE-10-ES*": Portal Estadístico del SNS, Ministerio de Sanidad, Servicios Sociales Igualdad. Accessible a: http://www.msssi.gob.es/estadEstudios/estadisticas/normalizacion/home.htm [consultat maig, 2016]

[18] Goetz R. "*What is the fundamental difference between "ETL" and "ELT" in the world of big data?*"; IBM Data Warehousing; maig 2016. Accessible a: https://ibmdatawarehousing.wordpress.com/2015/05/13/goetz-etl-bigdata/ [consultat març, 2016]

[19] Hadoop 1.2.1 Documentation. "*MapReduce Tutorial*": apache.org; abril 2013. Accessible a: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html [consultat març, 2016]

[20] Dean J, Ghemawat S. "*MapReduce: Simplified Data Processing on Large Clusters*": OSDI 2004, Google. Accessible a: http://static.googleusercontent.com/media/research.google.com/es//archive/mapreduce-osdi04.pdf [consultat abril, 2016]

[21] Yahoo Developer Network. "*Module 4: MapReduce*": Yahoo. Accessible a: https://developer.yahoo.com/hadoop/tutorial/module4.html#basics [consultat abril, 2016]

[22] Sims S. "PLN: Procesamiento del lenguaje natural": desembre 2015. Accessible a https://businessanalyticsdata.wordpress.com/2015/12/21/pln-procesamiento-del-lenguaje-natural/ [consultat abril, 2016]

[23] HPCC SYSTEMS*. Accessible a https://hpccsystems.com/ [consultat abril, 2016]

[24] Disco. Accessible a: http://discoproject.org/ [consultat març, 2016]

[25] Hadoop. "*What Is Apache Hadoop?*" (disponible en pdf): Apache.org; darrera publicació febrer, 2016. Accessible a: https://hadoop.apache.org/ [consultat març, 2016]

[26] Markey SC. "*Deploy an OpenStack private cloud to a Hadoop MapReduce environment*": developerWorks, IBM; octubre 2012. Accessible a: http://www.ibm.com/developerworks/cloud/library/cl-openstack-deployhadoop/ [consultat març, 2016]

[27] Loughran S. "*Products that include Apache Hadoop or derivative works and Commercial Support*": Hadoop Wiki, Apache.org; darrera actualització desembre 2014. Accessible a: http://wiki.apache.org/hadoop/Distributions and Commercial Support [consultat març, 2016]

[28] Hausenblas M. "*Lambda Architecture*": Developer Central, MapR*; https://www.mapr.com/developercentral/lambda-architecture [consultat abril, 2016]

[29] Spark docs. "*Spark Overview*": Apache Spark*. Accessible a: http://spark.apache.org/docs/latest/ [consultat abril, 2016]

[30] "Apache Spark™ is a fast and general engine for large-scale data processing": Apache Spark*. Accessible a: http://spark.apache.org/ [consultat abril, 2016]

[31] "*Spark Streaming Programming Guide*": Apache Spark*. Accessible a: http://spark.apache.org/docs/latest/streaming-programming-guide.html [consultat abril, 2016]

[32] "*Concepts*": Apache Storm*. Accessible a: http://storm.apache.org/releases/2.0.0-SNAPSHOT/Concepts.html [consultat abril, 2016]

[33] "*Why use Storm?":* Apache Storm*. Accessible a: http://storm.apache.org/index.html [consultat abril, 2016]

[34] MapR 5.0 Documentation. "*Running Storm on a MapR Cluster*": MapR*. Accessible a: http://doc.mapr.com/pages/viewpage.action?pageId=28213843 [consultat abril, 2016]

[35] "*List Of NoSQL Databases*". Accessible a http://nosql-database.org/ [consultat marc, 2016]

[36] Siemens Healthineers. "*Trend Chapter 4: Benefiting from Big Data*": Thinking Healthcare Ahead: Infographics for download, Siemens; 2015. Accessible a: https://www.healthcare.siemens.com/news-and-events/mso-infographics-download [consultat abril, 2016]

[37] Feldman B, Martin EM, Skotnes T. "*Big Data in Healthcare. Hype and Hope*": Consultora Dr. Boonie 360º, octubre 2012. Accessible a: https://es.scribd.com/doc/107279699/Big-Data-in-Healthcare-Hype-and-Hope [consultat maig, 2016]

[38] Raghupathi W, Raghupathi V. *"Big data analytics in healthcare: promise and potential*": BioMed Central. Health Information Science and Systems 2014 2:3. doi: 10.1186/2047-2501-2-3; febrer 2014. Accessible a: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/ [consultat abril, 2014]

[39] García J. "*La medicina del futuro pasa por Big Data*"; A un clic de las TIC, Telefónica; octubre 2014. http://aunclicdelastic.blogthinkbig.com/la-medicina-del-futuro-pasa-por-big-data/ [consultat abril, 2016]

[40] García J, García R, Hdez. de Rojas F, López V, Nuñez M. "*Big data. El poder de convertir datos en decisiones"*: A un clic de las TIC, Telefónica. Madrid, gener 2016. Accessible a: http://www.aunclicdelastic.com/wp-content/uploads/eBook-BIG-DATA-AunClicdelasTIC.pdf [consultat abril, 2016]

[41] Sánchez JJ. *"La privacidad de los datos de salud en la era digital"*: A un clic de las TIC, Telefónica; març 2015. Accessible a: http://aunclicdelastic.blogthinkbig.com/la-privacidad-de-los-datos-de-salud-en-la-era-digital/ [consultat maig, 2016]

[42] Byrd TA, Harbert RJ, Kung L, Ting C, Wang Y. "*Beyond a Technical Perspective: Understanding Big Data Capabilities in Health Care*": 2015 48th Hawaii International Conference on System Sciences, IEEE Computer Society, 3044-3053, DOI 10.1109/HICSS.2015.368. Accessible a: https://www.computer.org/csdl/proceedings/hicss/2015/7367/00/7367d044.pdf [consultat maig, 2016]

[43] IDC press Release. "EMEA Big Data Datacenter Infrastructure Market to Triple by 2019, Reaching $5.4B, Says IDC": IDC; gener 2016. Accessible a: https://www.idc.com/getdoc.jsp?containerId=prEMEA40935916 [consultat abril, 2016]

[44] Mell P, Grance T. "*The NIST Definition of Cloud Computing*": Special Publication 800-145, National Institute of Standars and Technology. Gaithersburg, MD 20899-8930; setembre 2011. Accessible a: http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf [consultat març , 2016]

[45] Belle A. "Construir una Cloud híbrida: La TI como servicio": IDC; Octubre, 2015. Disponible a: http://www.acens.com/blog/wp-content/images/construir-cloud-hibrida-idc-emc-informe-blog-acens-cloud.pdf [consultat abril, 2016]

[46] Intel IT Center. "*Big Data in the Cloud: Converging Technologies*": Intel; 2015. Accessible a: http://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/big-data-cloud-technologies-brief.pdf [consultat març, 2016]

[47] ITU-T Recommendations. *ITU-T Y.3600*: International Telecommunication Union; novembre 2015. Accessible a: http://www.itu.int/itu-t/recommendations/rec.aspx?rec=12584 [consultat març, 2016]

[48] Marr B. Big Data-As-A-Service Is Next Big Thing": Forbes; abril 2015. Accessible a: http://www.forbes.com/sites/bernardmarr/2015/04/27/big-data-as-a-service-is-next-big-thing/ [consultat abril, 2016]

[49] Morejon E. "Big Data as a service is the next Big Thing": febrer 2016. Accessible a: https://businessanalyticsdata.wordpress.com/category/bdaas/ [consultat abril, 2016]

[50] Tibco* Spotfire*, Accessible a: http://spotfire.tibco.com/solutions/technology/big-data [consultat maig, 2016]

[51] Cazena. Accessible a: http://www.cazena.com [consultat maig, 2016]

[52] Qubole Data Service*. Accessible a: https://www.qubole.com [consultat maig, 2016]

[53] Doopex. Accessible a: https://doopex.com [consultat maig, 2016]

[54] Fernandes J. "*Hortonworks Data Platform and OpenShift: Combining Big Data and Rapid App Development*"; OpenShift, RedHat; abril 2014. Accessible a: https://blog.openshift.com/combining-big-data-and-rapid-application-development-openshift-and-hortonworks-data-platform/ [consultat maig, 2016]

[55] Tibco* Jaspersoft*. Accessible a: https://www.jaspersoft.com/cloud-analytics [consultat maig, 2015)

[56] 23andMe. Accessible a: https://www.23andme.com/en-int/ [consultat maig, 2016]

[57] Proscia, Accessible a : https://proscia.com/documents/PathologyCloudInfrastructure.pdf [consultat maig, 2016]

[58] Proscia Patology Cloud*. Accessible a: https://proscia.com/platform [consultat maig, 2016]

[59] PatientsLikeMe*. Accessible a: https://www.patientslikeme.com/ [consultat maig, 2016]

[60] Fifbit. Accessible a: https://www.fitbit.com/es [cunsultat maig, 2016]

[61] Apple Watch. Accessible a: http://www.apple.com/es/watch/health/ [consultat maig, 2016]

[62] Topol, E. "Top 10 Avances Tecnológicos 2015 según Eric Topol": Diagnostrum; gener 2016. Accessible a: http://blog.diagnostrum.com/2016/01/08/top-10-avances-tecnologicos-2015-segun-eric-topol/

[63] Comstock J. " *Dexcom CGM app is available, unofficially, for Android Wear users*": mobihealth news; maig 2015. accessible a: http://mobihealthnews.com/43331/dexcom-cgm-app-is-available-unofficially-for-android-wear-users [consultat maig, 2016]

[64] Lumify de Philips. Accessible a: https://www.lumify.philips.com/web/ [consultat maig, 2016]

[65] PillCam. Accessible a: http://pillcamcolon.com/ [consultat maig, 2016]

[66] Public Healt. "*Medical intelligence in Europe*": European Commission; Disponible a: http://ec.europa.eu/health/preparedness_response/generic_preparedness/planning/medical_intelligence_en.htm. [consultat abril, 2016]

[67] Propeller Healt, Accessible a: https://www.propellerhealth.com/ [consultat maig, 2016]

[68] IBM News. "*IBM crea una nueva unidad de negocio, Watson Health, orientada al sector sanitario*": IBM (NYSE: IBM); Madrid, abril 2015. Accessibles a: https://www-03.ibm.com/press/es/es/pressrelease/46621.wss [consultat, maig 2016]

[69] IBM News. "*IBM Watson Health anuncia nuevos acuerdos y servicios en la nube, así como su nueva sede*": Massachusetts, EEUU, IBM (NYSE:IBM), setembre 2015. Accessible a: http://www-03.ibm.com/press/es/es/pressrelease/47648.wss [consultat maig, 2016]

[70] IBM Care Management. Accessible a: http://www-03.ibm.com/software/products/es/IBM-care-management [consultat abril, 2016]

Altes fonts consultades:

Gómez JL, Conesa i Caralt J. *"Introducción al big data" Barcelona"*: FUOC, PID_00209840, Universitat Oberta de Catalunya; juny 2015

Miranda A. "Big Intelligence": Fundacion EOI, ISBN 978-84-15061-61-8; Madrid, 2015. Accessible a: http://a.eoi.es/bigintelligence [consultat març, 2016]

Torres V, "*Del cloud computing al big data*": FUOC, PID_00194203, Primera Edició, Eureca Media, SL, CC-BY-NC-N; Barcelona, setembre 2012. Accessible a: http://www.jorditorres.org/wp-content/uploads/2012/03/Del.Cloud_.Computing.al_.Big_.Data_.JordiTorres.ES_.pdf [consultat març, 2016]

Coulouris G, Dollimore J, Kindberg T, Blair G. "*Distributed Systems: Concepts and Design"* (5th Edition) ISBN 978-0-13-214301-1: Addison-Wesley, Boston, 2012.