
Fonaments de *big data*: arquitectura del sistema

PID_00264329

José López Vicario

Temps mínim de dedicació recomanat: 2 hores



José López Vicario

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Maria Pujol Jover (2019)

Primera edició: març 2019
© José López Vicario
Tots els drets reservats
© d'aquesta edició, FUOC, 2019
Av. Tibidabo, 39-43, 08035 Barcelona
Disseny: Manel Andreu
Realització editorial: Oberta UOC Publishing, SL

Cap part d'aquesta publicació, incloent-hi el disseny general i la coberta, no pot ser copiada, reproduïda, emmagatzemada o transmesa de cap manera ni per cap mitjà, tant si és elèctric com químic, mecànic, òptic, de gravació, de fotocòpia o per altres mètodes, sense l'autorització prèvia per escrit dels titulars del copyright.

Índex

Introducció	5
1. Motivació de les dades massives en un entorn industrial ...	6
2. Arquitectura d'un sistema <i>big data</i>	8
3. Fonts i tipus de dades en l'empresa industrial	9
4. Extracció i homogeneïtzació de dades	10
4.1. Importància de la qualitat de les dades	10
4.2. Dades brutes i dades processades	10
4.3. Mètodes d'extracció de dades	12
5. Computació i emmagatzematge al núvol	14
6. Eines per a l'anàlisi de dades	16
7. Exemples d'aplicació	17
Resum	18
Bibliografia	19

Introducció

L'objectiu d'aquest material és presentar l'arquitectura d'un sistema *big data* i les consideracions que s'han de tenir en compte quan s'aplica a un entorn industrial.

En primer lloc, es motiva l'ús de la tecnologia *big data* en la indústria i es presenten els elements principals que solen formar part d'una arquitectura *big data*, des de l'adquisició fins a l'emmagatzematge de dades i la seva posterior anàlisi.

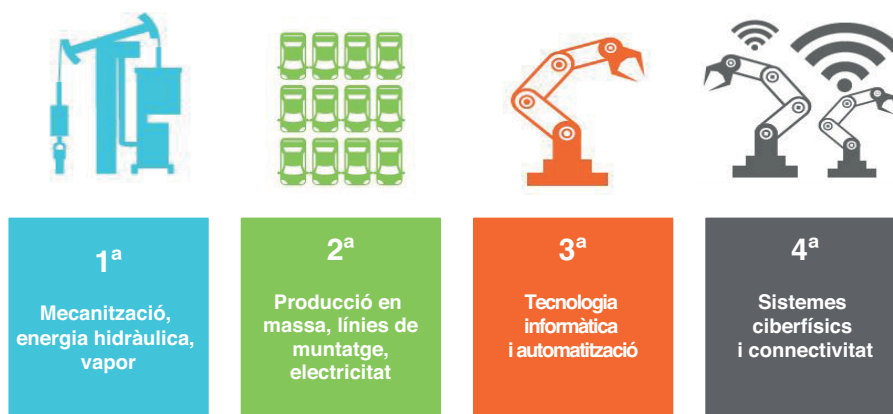
En segon lloc, es fa una revisió de la naturalesa de les fonts de dades principals que es poden trobar en l'entorn industrial i es presenten les consideracions que cal tenir en compte a l'hora de reunir aquestes dades. En concret, es fa èmfasi en la importància de tractar dades rellevants, és a dir, a treballar amb dades de qualitat i en els mètodes existents per extreure-les.

Finalment, s'aborda el concepte d'emmagatzematge en núvol (*cloud storage*) de les dades recopilades i la posterior computació d'aquestes dades per extreure'n els patrons d'interès. També s'exposen les principals eines disponibles, tant lliures com propietàries, per dur a terme anàlisis de dades i, a tall d'exemples d'aplicació, es presenten un conjunt d'aplicacions totalment alineades amb l'entorn industrial.

1. Motivació de les dades massives en un entorn industrial

Actualment, la indústria viu una quarta revolució industrial, coneguda com a **indústria 4.0** (vegeu la figura 1 per a un resum de les quatre revolucions industrials). Aquesta nova revolució es basa, principalment, en la connectivitat que ens ofereixen els sistemes ciberfísics (gràcies a l'auge de les xarxes de sensors i actuadors sense fil o bé al paradigma IoT, és a dir, a l'internet de les coses), la qual cosa permet prendre encara més mesures en els processos industrials i, al seu torn, proporcionar sistemes d'actuació més intel·ligents i en temps real. Perquè això sigui possible s'han d'explotar altres tecnologies que actualment també estan en alça, com ara la informàtica en núvol (*cloud computing*) i les dades massives (*big data*). La primera permet emmagatzemar dades d'una manera més eficient: amb un cost menor i amb accessibilitat des de qualsevol lloc. Les tecnologies *big data*, d'altra banda, possibiliten analitzar conjunts massius de dades. L'objectiu d'aquest material és mostrar com es planteja aquesta última tecnologia en l'entorn industrial.

Figura 1. Evolució industrial



Quan es fa referència a *big data*, es parla de les quatre V, és a dir, del volum, de la velocitat, de la varietat (que poden tenir les dades que s'estan tractant) i de la veracitat (pel que fa al problema de la veracitat de dades). Aplicat a l'entorn industrial, el concepte *big data* fa referència a la gran quantitat de dades generades per un conjunt de fonts molt diferents relacionades amb aquest entorn. Aquestes fonts de dades poden ser sensors de la maquinària, consum energètic, preu de les matèries primeres, planificació de processos i temps d'execució, utilització de recursos, etc.

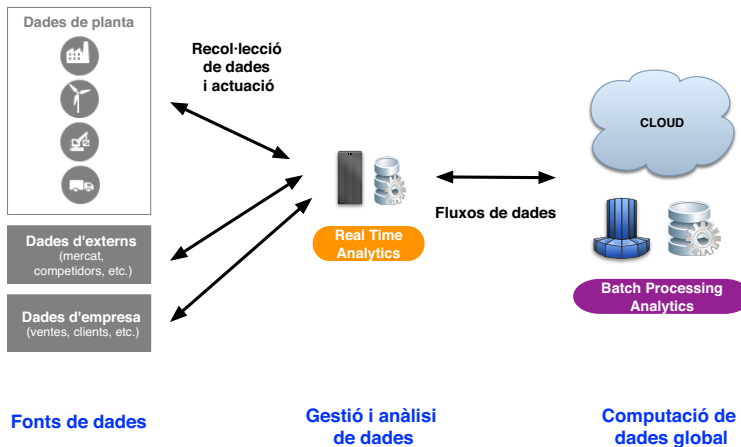
Com passa en altres entorns, la motivació de la utilització de la tecnologia *big data* és donada pels beneficis, en un àmbit econòmic, que es poden obtenir a l'hora d'analitzar les dades, tant reduint costos com incrementant vendes. En l'entorn industrial, aquests beneficis es poden obtenir mitjançant la gestió eficient dels recursos industrials, l'extracció de patrons per conèixer o millorar processos i la predicció de comportaments per adaptar-se a demandes de producció variants o prevenir anomalies (és a dir, per millorar la gestió de les operacions). Tot i que en l'apartat 7 es presentaran amb més detall exemples d'aplicacions de les *big data* en una planta industrial, vegem-ne algunes: optimització del procés de producció augmentant la seva capacitat mitjançant la detecció i l'optimització de colls d'ampolla; manteniment predictiu basat en l'ús de l'històric de dades de sensors per predir quan és necessari reparar o substituir maquinària; producció eficient energèticament a partir de l'anàlisi de dades de consum i la seva relació amb els tipus de productes fabricats.

En un àmbit més global, és a dir, en el cas d'una companyia amb diferents plantes repartides en un àmbit geogràfic, l'anàlisi de les dades es pot fer servir per millorar el rendiment de les plantes de producció i la interacció entre elles (és a dir, per millorar-ne l'estratègia en operacions). Per exemple, si es comparen els resultats entre plantes, és possible detectar bones pràctiques aplicables a les plantes més ineficients, millorar la flexibilitat en un àmbit de producció global i reduir els costos totals de producció (analitzant costos laborals, de transport o d'adquisició de matèries primeres o energia a les diferents regions).

2. Arquitectura d'un sistema *big data*

A la figura 2 es presenta l'esquema que es fa servir habitualment per implementar un sistema *big data* en l'entorn industrial.

Figura 2. Arquitectura d'un sistema *big data* en un entorn industrial



Tal com es pot observar, aquest esquema està format per:

- **Fonts de dades:** elements, dispositius o generadors de dades utilitzats en l'anàlisi duta a terme pel sistema *big data*. Com detallem en l'apartat següent, l'entorn industrial presenta un conjunt molt heterogeni pel que fa al tipus de fonts.
- **Gestió i anàlisi de dades:** conjunt de mecanismes que duen a terme tasques de computació i actuació en un àmbit de planta industrial. S'encarreguen de preprocessar, netejar i preparar les dades per poder emmagatzemar-les en servidors d'una manera més organitzada, així com d'executar anàlisis preliminars per a aquelles accions que requereixen actuació en temps real.
- **Computació de dades global:** en aquesta etapa, les dades emmagatzemades en servidors (propietaris o serveis al núvol) s'analitzen amb tècniques d'anàlisi de dades que obtenen millors prestacions, ja que les dades s'analitzen com un conjunt i sense restriccions temporals.

3. Fonts i tipus de dades en l'empresa industrial

En un entorn industrial, hi ha un conjunt molt ampli de dades. A continuació, es presenten les principals fonts de dades:

- **Dades de mercat:** dades relacionades amb el mercat o els mercats que l'empresa està abordant (per exemple, el mercat automobilístic dels Estats Units). En aquest context, algunes de les dades rellevants serien: la grandària del mercat (normalment és mesurada en euros, dòlars o la unitat monetària pertinent), la quota de mercat (percentatge del mercat capturat per empresa), les dades sobre competidors, la tendència de la grandària del mercat, etc.
- **Dades de proveïdors:** en aquest cas, les dades es refereixen als subministradors de matèries primeres utilitzades per l'empresa per generar els seus productes o serveis; és a dir, dades relacionades amb els proveïdors, la seva naturalesa (mida, poder de negociació), la quantitat, els preus oferts, el temps de lliurament i la qualitat del servei.
- **Dades de logística:** dades relacionades amb el transport, l'emmagatzematge i la distribució de les mercaderies als clients. Segons el tipus de sector al qual pertany l'empresa, aquestes dades poden ser essencials per millorar la competitivitat. Això és així en sectors en els quals es treballa amb un inventari elevat, ja que l'inventari i el seu transport o emmagatzematge s'han de finançar fins que es duu a terme la venda.
- **Dades de planta:** en aquest cas, el concepte fa referència a un conjunt més ric de fonts de dades, tenint en compte el context d'indústria 4.0 i l'ús de IoT per poder prendre més mesures de baix cost. Aquestes mesures abastarien dades de sensors equipats a la maquinària de la planta, dades sobre consum energètic, dades sobre la capacitat/*throughput*/retard en els processos de producció o dades sobre la planificació de la producció.
- **Dades de vendes:** dades relacionades amb les vendes efectuades per l'empresa, la seva tendència, les vendes per tipus de producte o servei, el mercat o el client.
- **Dades de clients:** informació rellevant sobre els clients de l'empresa, la seva grandària, el sector al qual pertanyen, el seu poder de negociació, les vendes anuals, els preus i els productes oferts, la seva localització geogràfica, el perfil que tenen i el segment al qual pertanyen.
- **Dades rellevants de l'entorn:** dades relatives al context que envolta l'activitat empresarial, com ara els països en què s'opera, la legislació i la normativa aplicables, els costums i les característiques de la regió.

4. Extracció i homogeneïtzació de dades

4.1. Importància de la qualitat de les dades

Tal com s'ha observat en l'apartat anterior, les dades utilitzades en l'entorn industrial provenen de fonts heterogènies i són de naturalesa molt diversa. Això fa que disposar d'un conjunt de dades organitzades i ben estructurades que puguin ser analitzades posteriorment resulti una tasca molt difícil. Per aquest motiu, en aquest entorn és vital l'ús de tècniques de preprocessament que serveixin per organitzar totes les dades i garantir que siguin completes, contínues i sincronitzades.

Així mateix, en l'entorn industrial els atributs analitzats tenen una clara interpretació física, i garantir la integritat de les dades que els caracteritzen també és molt important per no dur a terme anàlisis o decisions errònies. Disposar de dades de baixa qualitat (alterades per sorolls) o mal emmagatzemades (per errors de sensors, mala connectivitat o errors d'emmagatzematge) pot portar a caracteritzar erròniament les relacions causa-efecte i provocar de manera catastròfica la presa de decisions basada en l'anàlisi de les dades.

Com a conclusió, en aquest apartat cal esmentar que el coneixement sobre l'entorn industrial és fonamental per poder desenvolupar projectes *big data*. Aquest coneixement, així com el fet de tenir un equip apropiat, permet que sigui possible dissenyar tècniques de preprocessament adequades i detectar quan els resultats obtinguts no són fiables perquè les dades han estat de baixa qualitat.

4.2. Dades brutes i dades processades

Quan es parla de **dades brutes** es fa referència a les dades rebudes directament de les seves fonts. Aquestes dades estan caracteritzades per les propietats següents:

- inexistència de format unificat
- possibilitat de valors no assignats (*missing values*)
- possibilitat de mesures corruptes a causa del soroll o errors de comunicació o d'emmagatzematge de les dades
- dificultat de ser tractades directament per algorismes d'anàlisi de dades

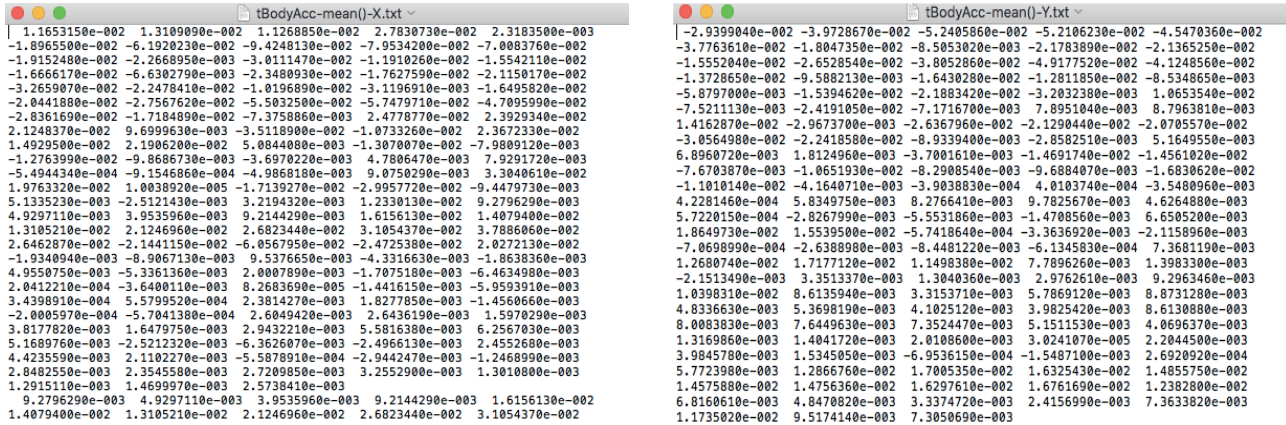
Pel que fa a les **dades processades**, són aquelles a les quals s'ha aplicat algun dels tipus de tractament esmentats en l'apartat 2: preprocessament, neteja o preparació de dades. L'objectiu d'aquest tractament és que les dades puguin ser emmagatzemades en unes estructures de dades, generalment anomenades *data sets*, per analitzar-les després. És a dir, les dades poden expressar-se en forma de taula: cada variable o atribut és en una columna, cada valor de cada variable se situa en una fila i cada columna té el nom de la variable pertinent.

En les figures 3 i 4 es presenta de manera il·lustrativa com es vol que s'organitzin les dades per poder ser analitzades, en contraposició a com se solen trobar normalment a la pràctica. Observeu com en el cas de les dades processades (figura 3), les dades obtingudes d'un conjunt d'acceleròmetres (per exemple, equipats en un robot industrial), es presenten ben organitzades, homogeneïtzades i preparades per executar mecanismes d'anàlisi de dades. En el segon cas, en canvi, es mostren les dades brutes obtingudes de diferents sensors. En aquest cas, les dades presenten un aspecte més caòtic, ja que s'obtenen consultant directament la font que les genera. Per tant, s'observa la necessitat de manipular-les i processar-les per poder generar el *data set* desitjat.

Figura 3. Dades processades de mesures d'acceleròmetres equipats en un robot industrial

	subject	activity	tBodyAcc-mean()-X	tBodyAcc-mean()-Y	tBodyAcc-mean()-Z	tGravityAcc-mean()-X	tGravityAcc-mean()-Y	tGravityAcc-mean()-Z	tBodyAccJerk-mean()-X
1	1	andar	0.2773308	-0.017383819	-0.11114810	0.9352232	-0.282165021	-0.068102864	0.07404163
31	1	bajarEsc	0.2891883	-0.009918505	-0.10756619	0.9318744	-0.266610339	-0.062119959	0.05415532
61	1	depie	0.2789176	-0.016137590	-0.11060182	0.9429520	-0.272983832	0.013490582	0.07537665
91	1	sentado	0.2612376	-0.001308288	-0.10454418	0.8315099	0.204411593	0.332043703	0.07748252
121	1	subirEsc	0.2554617	-0.023953149	-0.09730200	0.8933511	-0.362153364	-0.075402940	0.10137273
151	1	tumbado	0.2215982	-0.040513953	-0.11320355	-0.2488818	0.705549773	0.445817720	0.08108653
2	2	andar	0.2764266	-0.018594920	-0.10550036	0.9130173	-0.346607090	0.084727087	0.06180807
32	2	bajarEsc	0.2776153	-0.022661416	-0.11681294	0.8618313	-0.325780101	-0.043889016	0.11004062
62	2	depie	0.2779115	-0.018420827	-0.10590854	0.8969286	-0.370062697	0.129747161	0.07475886
92	2	sentado	0.2770874	-0.015687994	-0.10921827	0.9404773	-0.105630024	0.198726769	0.07225644
122	2	subirEsc	0.2471648	-0.021412113	-0.15251390	0.7907174	-0.416214890	-0.195888239	0.07445078
152	2	tumbado	0.2813734	-0.018158740	-0.10724561	-0.5097542	0.752536639	0.646834880	0.08259725
3	3	andar	0.2755675	-0.017176784	-0.11267486	0.9365067	-0.261986358	-0.138107866	0.08147459
33	3	bajarEsc	0.2924235	-0.019355408	-0.11613984	0.9390578	-0.228829214	-0.102352758	0.07256893
63	3	depie	0.2800465	-0.014337656	-0.10162172	0.9350308	-0.301735130	0.024763107	0.07509006
93	3	sentado	0.2571976	-0.003502998	-0.09835792	0.9010990	0.127303379	0.139020582	0.07260984
123	3	subirEsc	0.2608199	-0.032410941	-0.11006486	0.8835334	-0.382851154	-0.162943955	0.04268810
153	3	tumbado	0.2755169	-0.018955679	-0.10130048	-0.2417585	0.837032100	0.488703185	0.07698111
4	4	andar	0.2785820	-0.014839948	-0.11140306	0.9639997	-0.085854027	0.127764113	0.07835291

Figura 4. Dades brutes de diferents acceleròmetres



Finalment, la taula següent mostra un resum de la comparació entre dades brutes i dades processades.

Taula 1

Dades brutes	Dades processades
Dades extretes directament de la font original.	Dades organitzades i a punt per ser analitzades.
Difícilment poden utilitzar-se per dur a terme un procés d'anàlisi.	Pot haver-hi unes normes més o menys estrictes per analitzar-les (legals, tecnològiques, estadístiques).
L'anàlisi de les dades requereix un processament previ.	L'anàlisi pot requerir algunes operacions prèvies (aglutinació, fusió, segmentació).
Poden requerir una o més transformacions.	Cal documentar tots els passos de l'anàlisi.

4.3. Mètodes d'extracció de dades

Tal com s'ha presentat en l'apartat 3, en l'entorn industrial els conjunts de dades són molt heterogenis. Per aquest motiu, es requereix l'ús de diferents mètodes d'extracció de dades segons la font de dades. Les eines de programari utilitzades per a l'anàlisi de dades (vegeu l'apartat 6) disposen de multitud de llibreries i funcionalitats per poder extreure dades. El desenvolupador tan sols hauria de consultar cada cas específic segons el seu entorn de treball. Alguns casos usals serien els següents:

- **Accés a bases de dades:** per extreure dades sobre clients, proveïdors, informació de l'empresa, vendes, preus, etc.
- **Lectura de fitxers locals:** per extreure informació similar a la del punt anterior, però que ha estat emmagatzemada en fitxers Excel, per exemple.
- **Descàrrega de fitxers del web:** per extreure dades disponibles en pàgines web amb informació sobre l'entorn empresarial.

- **Lectura de formats web:** per extreure informació continguda en formats habitualment utilitzats al web, com ara XML o JSON.
- **Web scrapping:** fa referència a l'extracció directa de les dades contingudes en una pàgina web. L'aplicació industrial seria similar als dos casos anteriors.
- **Lectura de dades recollides per sensors:** normalment es duu a terme, en un entorn industrial, mitjançant plataformes basades en SCADA o l'ús de nodes recollidors (anomenats *gateways*), encarregats d'agregar les mesures de diferents sensors.

5. Computació i emmagatzematge al núvol

Com s'ha esmentat en l'apartat 1, una de les tecnologies explotades en el paradigma d'indústria 4.0 és la tecnologia *cloud* o núvol. Aquesta tecnologia es basa en un nou concepte, en el qual proveïdors de serveis ofereixen serveis d'emmagatzematge o computació per internet. És a dir, els clients d'aquests serveis utilitzen les infraestructures de servidors d'emmagatzematge i computació dels proveïdors per emmagatzemar les seves dades o dur a terme tasques de processament.

En l'àmbit de *big data*, aquesta tecnologia facilita la possibilitat d'operar amb un gran volum de dades. Això es deu al fet que el principal avantatge que ofereixen les solucions d'emmagatzematge basades en el núvol és que el seu cost se sol basar en un pagament per ús de recursos d'emmagatzematge (per exemple, un pagament basat en euros/GB). Per tant, l'empresa que opta per aquesta opció no ha d'incórrer en una gran inversió en infraestructura (maquinari o servidors) i el seu posterior manteniment (de prop del 30% en alguns casos) necessari per emmagatzemar les *big data* (Minelli i altres [2012]; Minelli, Chambers i Dhiraj).

Així doncs, som davant un model basat en OPEX (despeses operatives, és a dir, despeses mensuals o anuals per un servei) diferent del tradicional, basat en CAPEX (inversions en béns de capitals, les quals requereixen una certa capacitat financera). El model basat en OPEX ha anat guanyant molta popularitat en els últims anys, a causa de la crisi financera viscuda. També és un model molt adequat per a empreses que no són gaire grans. Un altre avantatge és que la capacitat d'emmagatzematge oferta pel proveïdor pot escalar segons les necessitats del client i que el proveïdor actualitza la plataforma automàticament. En un model CAPEX, en canvi, caldria anar reinvertint en noves infraestructures si el volum de les dades anés creixent.

No obstant això, encara hi ha moltes empreses que no han migrat al núvol a causa de la desconfiança que pot generar tenir les dades emmagatzemades en instal·lacions de tercers. En aquest sentit, cal esmentar que els principals proveïdors, com Amazon (AWS, Amazon Web Services), garanteixen el nivell dels serveis que ofereixen mitjançant SLA (*service level agreement*). Així mateix, els nivells d'inversió i especialització d'aquests proveïdors permeten oferir més seguretat que la que moltes empreses podrien tenir en les seves pròpies instal·lacions.

A part de serveis d'emmagatzematge, molts proveïdors de núvol ofereixen serveis de computació o de connectivitat de dispositius. Seguint amb l'exemple d'Amazon, aquest proveïdor ofereix serveis com Amazon EC2 (capacitat de

còmput variable) o AWS IoT (plataforma que permet la connectivitat de dispositius al núvol). De nou, són serveis basats en pagament per consum (sota demanda o reservat), interessants per a aquelles empreses que no disposin de capacitat financera per abordar una gran inversió.

6. Eines per a l'anàlisi de dades

Actualment, el conjunt d'eines per a l'anàlisi de dades és ampli. En un àmbit d'eines propietàries trobem solucions dels principals proveïdors, com ara General Electrics (<https://www.ge.com/digital/>, amb productes com Asset Performance Management o Predix); SAS (https://www.sas.com/es_es/home.html, amb productes i solucions com Supply Chain Intelligence o Customer Intelligence); NTT Data (<https://us.nttdata.com/en/Industries>) o també l'empresa IBM (<https://ibm.co/2zSnkG9>).

No obstant això, l'ús d'eines lliures s'ha fet molt popular en els últims anys a causa del suport que han rebut de nombroses comunitats i per les excel·lents prestacions que ofereixen. Dins del gran ecosistema d'eines *big data* lliures, en aquest apartat en destaquem un conjunt:

- **Recol·lecció de dades:** eines encarregades de recollir dades emeses per fonts i d'organitzar-les posteriorment per assegurar que no es perden. Solucions com Apache Flume, Apache Sqoop o Apache Kafa.
- **Emmagatzematge de dades:** eines encarregades d'emmagatzemar dades massives perquè posteriorment puguin ser analitzades. Solucions com HDFS o HBase.
- **Anàlisi de dades basades en *batch processing*:** eines encarregades d'analitzar dades emmagatzemades en bases de dades distribuïdes en servidors propis o en el núvol. Solucions com Hadoop o Apache Spark.
- **Anàlisi de dades en *stream processing*:** eines encarregades d'analitzar dades en temps real. Solucions com Apache Spark Streaming o Apache Storm.
- **Anàlisi estadística:** eines encarregades de dur a terme anàlisis estadístiques de les dades o d'aplicar algorismes *machine learning*. Solen estar basades en llenguatges de programació que es poden fer servir per desenvolupar sistemes executats en els dos punts anteriors (*batch* o *stream processing*). Solucions com R o Python.

7. Exemples d'aplicació

En la taula següent es presenten una sèrie d'exemples d'aplicació de les *big data* en el sector industrial.

Taula 2

Àmbit	Aplicació
Excel·lència operativa	<ul style="list-style-type: none"> • Automatització de presa de decisions paramètrica • Optimització de processos de negoci • Optimització de fluxos de producció • Optimització i monitoratge en temps real • Optimització de l'eficiència energètica • Control i assegurement de la qualitat • <i>Data center management</i>
Eficiència de procés i producte	<ul style="list-style-type: none"> • Millora en el disseny de processos i productes amb dades d'ús • Anàlisi i predicció de la demanda • Detecció de patrons i <i>root cause analysis</i> • Intel·ligència de mercat • Manteniment predictiu i preventiu d'equipament, instal·lacions i estructures • Preus dinàmics • <i>Product lifecycle</i>
Gestió de riscos	<ul style="list-style-type: none"> • Assegurances • Fraus • Seguretat (videoanàlisi en temps real)
Productes i serveis intel·ligents	<ul style="list-style-type: none"> • Dada com a producte i dada com a servei (<i>data as a service</i>) • Monitoratge remot • Edificis intel·ligents • Fàbrica connectada • Intel·ligència ambiental • Gestió del coneixement tècnic (diferents fonts d'informació)
Coneixement del client	<ul style="list-style-type: none"> • <i>Customer support</i> • Anàlisi de sentiment • <i>Targeting</i> • Recomanació • Prevenció de l'abandonament • <i>Customer lifetime value</i> • Anàlisi del <i>product mix</i> • <i>Personal analytics</i>

Resum

Aquest material té per objectiu presentar els fonaments de les dades massives aplicades a un entorn industrial. En primer lloc, s'ha motivat l'ús d'aquesta tecnologia en la indústria per poder satisfer les necessitats de la quarta revolució que s'està vivint actualment i, a continuació, s'ha mostrat l'arquitectura d'un sistema *big data* típic. Tenint en compte el gran conjunt de dades que pot haver-hi en un entorn industrial, s'han presentat les fonts de dades existents, la necessitat d'homogeneïtzar-les perquè puguin ser analitzades i els mètodes d'extracció de dades més utilitzats.

Finalment, s'ha descrit el concepte *cloud computing*, un dels pilars per facilitar la penetració de les *big data* en la indústria actual. Tot seguit s'han presentat les eines que es fan servir normalment en l'anàlisi de dades, com també alguns exemples d'aplicació industrial.

Referència bibliogràfica

M. Trovati; R. Hill; A. Anjum; S. Ying Zhu; L. Liu (2016). *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications*. Springer.

Bibliografia

Minelli, M.; Chambers, M.; Dhiraj, A. (2012). *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*. John Wiley and Sons.

Ministeri d'Indústria, Energia i Turisme (2016). *La transformación digital de la Industria española*. <<http://www.industriaconectada40.gob.es>>

Trovati, M.; Hill, R.; Anjum, A.; Ying Zhu, S.; Liu, L. (2016). *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications*. Springer.

