Anàlisi i aplicació del Big Data a petites empreses

Laia Aguilar Muñoz

Resum — En aquesta era de la informació, on la generació de dades és una *condició sine qua non*, sorgeix la necessitat de sentir que podem controlar tots i cadascun dels nostres moviments a la xarxa. Per això el Big Data, permet obtenir informació de les connexions entre dispositius i analitzar els registres detallats d'aquestes connexions. Sense una anàlisi posterior de totes aquestes bases de dades amb informació, sovint no estructurada, les dades recollides serien en va. En vista d'això, aquest treball valora i analitza diferents bases de dades i maneres en les quals s'emmagatzema la informació, juntament amb la comparativa de les eines de codi obert que una empresa pot emprar per treure partit a la informació, sense la necessitat d'una gran inversió econòmica en alternatives privades i per fer iniciar-se en l'extracció i l'analisis de les dades.

Paraules clau—Big Data, Hadoop, Spark, Base de dades, empresa, Apache

Abstract — We are in an information era, where collecting data is a pressing matter but also the feeling that we can control the amount of data we share online. Big Data gives access to information about connections between devices and analyzing the detailed record of each and every connection is now a possibility. However this information requires further analysis, to sort it out, otherwise the data is just collected in vain. With this problematic in mind, this paper analyzes and evaluates the different database and the ways in which information is stored. What's more, it compares the different open source tools a business can use to take advantage of this information, without necessarily investing in private solutions.

Index Terms— Big Data, Hadoop, Spark, Database, business, Apache

1 Introducció

DESPRÉS de passar anys, escollint la informació rellevant per emmagatzemar i així no acumular munts de papers, documents i arxivadors amb dades, tant personals com empresarials o governamentals, va arribar el moment en què aquest concepte va ser tergiversat gràcies a Internet i el creixement exponencial de dades que provocava aquest esdeveniment.

Amb l'evolució d'Internet i dels usuaris que navegaven i naveguen en ell, les arquitectures dissenyades inicialment i els motors de cerca que hi havia, no permetien l'abast a tot el que es produïa, ni permetia tenir constància del que passava i de les modificacions que hi havia en cada una de les dades emmagatzemades.

Gràcies a l'avanç de les diferents arquitectures i dels objectius fixats, va aparèixer el concepte de Big Data[1], per donar solució a conceptes obsolets i a canviar la forma de tractar grans volums de dades, per analitzar i transformar la informació en allò útil per a algú o alguna cosa.

És en aquesta línia, que aquest treball pretén donar resposta, a la necessitat que genera una empresa de recollir dades ininterrompudament i de forma fiable, permetent una millora en l'obtenció i el tractament de la informació, per tal que empreses i organitzacions puguin realitzar anàlisi o treure'n estadístiques i profit de tot allò que els usuari, sense adornar-se'n, proporcionen quan naveguen per Internet, quan envien notificacions d'errors, quan cerquen informació vinculada a organitzacions o en inesgotables situacions que els usuari ni hi pensen.

Al llarg del treball, es realitza l'estudi perquè les empreses pugui escollir un tipus de sistema, coneixent els diferents perfils i eines que hi ha al mercat, sabent-ne en més aprofundiment quins són els beneficis de cadascuna i els procediments que desenvolupen per a concloure en els resultats esperats i destacar les que més s'ajusten a les necessitats i possibilitats de l'organització.

Així com, a més de mostrar en marc teòric allò que s'ha de saber per començar a aplicar el Big Data, se'n fan les primeres configuracions d'una de les eines, que amb els mínim desenvolupament, es mostren resultats de l'eficàcia en la forma com el Big Data tracta les dades i el poc temps que necessita per mostrar-ne resultats.

- *E-mail de contacte: laia.aguilar@e-campus.uab.cat*
- *Mención realitzada: Tecnologies de la Informació.*
- *Treball tutoritzat per: Ramón Musach Pi (Dpt. d'Enginyeria de la Informació I de les Comunicacions, Universitat Autònoma de Barcelona).*
- *Curs 2016/17*

[1] Big Data, terme per definir conjunts de dades, que són tan gran o complex, que aplicacions de processament de dades tradicionals són inadequades per poder tractar-les. Té els reptes d'anàlisi, captura, recuperació de dades, recerca, intercanvi, emmagatzematge, transferència, visualització, consulta, actualització i privacitat d'informació. Sovint es refereix a l'ús d'analítica predictiva, analítica de comportament de l'usuari, o altres mètodes d'analítica de dades avançades que extreure'n valor a les dades.

2 OBJECTIUS

Aquest treball sorgeix de les inquietuds de saber-ne més en detall del que envolta al Big Data, a part dels articles periodístics que estan popularitzant el terme i de com els medi d'informació destaquen la seva importància en la societat, sense destacar el funcionament, i generant dubtes a organitzacions que volen aplicar-ho. En aquest treball es pretén plantejar un Big Data per petites empreses, amb tots els coneixements previs necessaris que això requereix pel primer contacte.

Quan es refereix a Big Data, s'entén per conjunt de dades que són complexes de processar, però per arribar fins a aquesta definició s'han de saber tractar les bases de dades, que són les antecessores i que en molts casos estan involucrades en el procés de Big Data. Per això, una de les finalitats és saber les diverses Bases de Dades que hi ha, per tal de tenir comparatives entre elles i aprofitar al màxim les dades en funció de la situació en que s'emprin, així com valorar els aspectes positius que poden proporcionar cadascuna i els inconvenients que podrien sorgir en la seva utilització.

Per tal de processar dades, fan falta eines que realitzin aquestes tècniques, cosa que genera la necessitat de conèixer les més utilitzades i quins són els aspectes d'aquestes eines que les fan destacar per sobre d'altres. I més concretament, perquè Hadoop és l'eina més utilitzada actualment i què la fa predominar per sobre les altres.

Esbrinar la manera de com es tracten, s'extreuen i es conclouen decisions a partir de les dades d'un Big Data. I conèixer les estratègies utilitzades, per tal de poder optimitzar la informació que els Big Data proporcionen i faciliten a les empreses.

Generalitzar els requeriments inicials i necessaris per aplicar Big Data i estandarditzar un procés, a partir de la informació obtinguda durant l'estudi del treball per empreses que no disposin d'aquests sistemes, per tal de treure informació de les dades que l'organització te recollides i no n'extreu tot el profit, així com tenir resposta en pocs segons de cerques de gran volum de dades.

3 ESTAT DE L'ART

Es pot considerar que l'ésser humà sempre ha fet un esforç per recopilar i analitzar les dades adquirides d'una manera o altra. Es podria remuntar a anys i anys aquesta ideologia i anar lligant caps fins arribar a l'objectiu, però no cal retrocedir tant per assolir una idea del Big Data.

Remuntant-se a la dècada dels 60, el que s'entén per Business Intelligence (BI) [2] va centrar l'anàlisi de les dades existents, en l'empresa i en estratègies per prendre decisions empresarials mitjançant l'ús de sistemes basats en fets de suport. A més, sorgien els primers intents i acostaments al emmagatzemament de dades modern juntament amb els primers centres de dades.

[2] Business Intelligence és l'ús de dades d'una empresa per facilitar la presa de decisions. Inclou tant la comprensió del funcionament actual de l'empresa com la predicció d'esdeveniments futurs, amb l'objectiu d'oferir coneixements per a donar suport a decisions empresarials.

A finals del segle XIX van començar a aparèixer conceptes importants, concretament a la dècada dels 90 va néixer el Internet de les Coses (IoT) on qualsevol cosa es podria mostrar en línia i pujar a internet per compartir amb altres persones. Conseqüentment al poc temps, el emmagatzemament digital va decaure fins al punt de ser més rentable que fer-ho amb paper, i arrel d'aquest successos, en aquesta mateixa dècada va néixer el motor de cerca més popular, Google.

Cap al 2000 sorgeixen les primeres idees del terme Big Data, en un treball acadèmic *Visually Exploring Gigabyte Datasets in Realtime (ACM)* [1] juntament amb les primeres definicions de "les tres V" del Big Data per donar sentit a Volum, Velocitat i Varietat que el terme representa, però no va ser fins l'any 2007 que es va consolidar el que actualment es coneix com a Big Data.

Des de llavors, la percepció de la generació i emmagatzemament de dades era més elevada que la que hi havia hagut fins als començaments de la civilització humana, cosa que despertava noves incògnites, sobre el tractat de grans volums de dades, de la privacitat, de la seguretat i de la propietat privada d'autors. El nivell de creixement era quasi incontrolable, fins ser al 2014 per primer cop, on la utilització del Internet mòbil superava l'ús del Internet als ordinadors d'escriptori, i per tant, més accessible per tots els entorns i no per un ús exclusius de feina.

Dia rere dia, la necessitat del maneig de les dades és un ingredient necessari i essencial per fer possible que la societat funcioni correctament, per això el terme Big Data segueix creixent i transformant-se per modificar el concepte d'analítica i de negoci en general.

Actualment el número de dispositius connectats ascendeix a dos vegades la població mundial [2], per tant l'economia impulsada per la vinculació a l'anàlisi d'aquests dispositius connectats pot afectar a molts sectors, inclús s'espera que el PIB mundial obtingui un impuls significatiu. Per tant, cada cop és més important el Big Data per a les empreses i que aquestes sàpiguen accedir a les eines, per crea oportunitats noves gràcies a l'anàlisi de gran volum de dades provinent de clients, proveïdors i competidors; i així descobrir estàndards, tendències i sentiments produïts, entre d'altres pel que el Big Data ofereix.

4 METODOLOGIA

S'han considerat diverses metodologies per desenvolupar el treball, però pel desenvolupament individual d'aquest projecte s'han descartat les que poden ser més utilitzades a empreses, ja que són principalment per aplicacions en col·lectiu o amb varis equips, cosa que no és el cas d'aquest treball, que es desenvoluparà de forma individual.

En el cas d'aquest projecte, es va considerar a l'inici aplicar el model de desenvolupament de software Cascada, que utilitza una metodologia d'iteracions i etapes definides que s'adapta molt bé a les necessitats. Gràcies a la seva linealitat com a metodologia de treball, quan s'acaba una fase es passa a la següent i així fins a obtenir el resultat marcat, ja que en aquest tipus de projecte es destaquen a l'inici els requisits i objectius que poden variar al llarg del desenvolupament però generalment s'ha d'assolir la finalitat preestablerta.

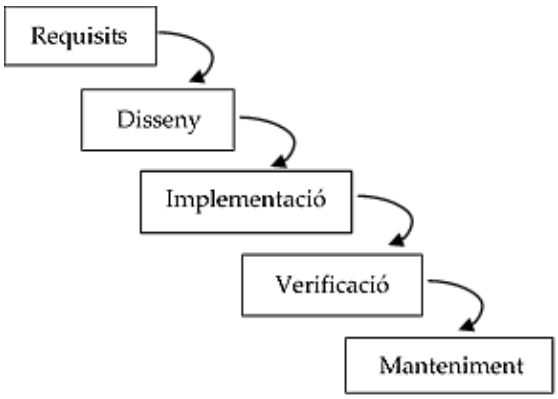


Fig. 1. Flux emprat per la metodologia Cascada

En aquest model, es defineixen les especificacions dels requisits, on a partir d'aquests requisits es fa un disseny de software previ per escollir la que es considera l'opció més encertada.

Les dues primeres fases serveixen per seguir amb la construcció del software i quan el software es pot considerar enllestit, s'integra amb tots els mòduls o altres grups de software que componen el software final.

Després d'aquestes fases, el model segueix amb les fase de validació i proves d'instal·lació del software i per últim, un cop acabat tot el procés inicial de les fases esmentades, s'hauria de fer un manteniment del software desenvolupat per no córrer riscos de desmilloraments o fallides.

Respecte als obstacles esperats del model Cascada al llarg de la redacció del treball, lamentablement s'ha comprovat que el fet d'haver de tancar etapes prèvies per poder continuar amb el correcte ritme de treball i poder avançar amb el desenvolupament d'altres etapes, ha dificultat més la tasca de realització. Ja que el fet de saber tots els detalls dels diferents sistemes plantejats no sempre és fàcil de trobar, ni es poden esmentar totes les característiques en un primer moment, perquè a mesura que la informació és més extensa i el domini del temari és més àgil, en surten més detalls rellevants que si s'haguessin sabut en una fase anterior s'hauria treballat millor, no obstant, es un bon mètode que agilitzar les tasques per saber quan s'ha de concloure una fase i començar-ne una altra.

4.1 Anàlisi de requeriments

Com a fase preliminar que promou la iniciativa del projecte és introduir l'aplicació del Big Data a petites empreses, perquè els grans negocis no han dubtat en la rendibilitat d'invertir recursos en sistemes i personal per implantar software i eines que permeten l'anàlisi i la gestió de grans volums de dades, però per a les pymes aquestes decisions d'inversió no són tan fàcils, ja que des d'un primer moment, es presenten com a solucions elevades per pressupostos reduïts.

Per tant, donar orientació i ensenyar a treure'n partit per desafiar el Big Data juntament amb una inversió mínima, són els requisits inicials amb que aquest treball s'inicia. Perquè petites iniciatives no necessàriament requereixen un software dissenyat exclusivament i propi per cada empresa, sinó que amb la configuració idònia d'eines de software lliure poden obtenir resultats iguals. Inclús empreses que dissenyen i executen estratègies en base a eines existents, obtenint millors rendiments que altres eines que ofereixen més complements però són menys ajustas als anàlisi de cada pyme dins dels diversos sectors.

5 DESENVOLUPAMENT

Per resoldre el plantejament dels Big Data a empreses, cal focalitzar i tenir constància d'allò que l'envolta pel correcte funcionament. El coneixement previ, és saber quin tipus de base de dades hi ha actualment per tal de poder escollir-ne la millor opció. I les eines que s'ajusten a les necessitats de les extraccions de dades i que permeten l'anàlisi aquestes.

5.1 Estudi de les diferents bases de dades

S'ha de plantejar quina és la diferencia entre base de dades i Big Data abans d'entrar en els detalls del Big data, ja que es podria considerar que les bases de dades són els antecedents d'aquesta nova tecnologia.

Quan es parla de base de dades s'entén que són dades netes i relacionades per tal de donar consistència a l'hora de ser classificades en els diversos contextos als que pertanyen, i sobretot que són dades íntegres d'alta qualitat, que mantenen un seguit de regles que han de complir i que permeten saber en tot moment les estructures que segueixen, sent aquesta definició la referència a base de dades SQL. Però aquestes base de dades tenen certes limitacions importants, ja que no són capaces d'afrontar dades no estructurades, tenint en compte la complexitat de les taules que cal dissenyar per tal de tenir totes les dades emmagatzemades i unides en estructures, així com la unió de diferents joins i funcions. Una altra limitació és l'escalabilitat, ja que no és el seu fort, perquè els usuaris que utilitzen aquests tipus dades han d'escalar les bases de dades relacionals en servidors potents, que són cars i difícils de manejar, a més de que han estar disponibles i encesos tots aquells servidors als quals les dades han d'accedir per adquirir alguna taula o informació que requereixi el disseny.

Per corregir aquests entrebancs o limitacions que oposa les base de dades SQL, en sorgeixen les anomenades no només SQL (Not Only SQL o NoSQL) que tracten esquemes de dades no estructurades, podent emmagatzemar menys dades en múltiples col·leccions i nodes sense requerir taules fixes. Aquest tipus de base de dades milloren l'escalabilitat, ja que amb les altres només hi havia escalabilitat vertical, és a dir, només es podien afegir recursos a un node en concret, com podia ser memòria. En canvi, amb aquest tipus de dades NoSQL, es té en compte l'escalabilitat horitzontal que, a més d'afegir memòria a un sol node, permet afegir més nodes al mateix i millorar el rendiment. El manteniment dels servidor NoSQL és menys costos, ja que són de codi obert i entre d'altres, la reparació automàtica és factible, i els models de dades són més simples i redueixen els requisits d'administració, sense la necessitat de requerir personal específic i capacitat per tasques concretes. També s'ha de

tenir en comte que suporta emmagatzematge en cache integrat, pel que augmenta el rendiment de sortida de les dades i el fa més àgil.

Però com passa amb moltes tecnologies, en surten millores que fan quedar endarrerides les que s'utilitzaven, fent que obtinguin limitacions amb les novetats. Encara que una de les característiques que es podria considerar positiva pel NoSQL és el fet de ser de codi obert, que alhora es podria convertir en una debilitat, perquè no hi ha estàndards definits ni bones pràctiques preestablertes per desenvolupar el codi, impedint que es trobin dues NoSQL iguals de referència i sent així, difícil trobar personal expert o desenvolupadors en aquesta tecnologia que conegui àmpliament el codi.

En canvi, Big data és una tecnologia, un framework de càlcul distribuït per processar grans volums de dades i per extreure'n patrons repetitius a partir de grans conjunts de dades, que proporciona velocitat, volum, varietat i veracitat a les dades. Generalment, aquest framework és de software de codi obert permetent que altres entitats utilitzin aquest codi per complimentar les necessitats pròpies i així, entre tots poder anar millorant els patrons d'extracció de dades.

Totes aquestes possibilitats que proporciona el Big data serveixen per explotar les dades que, mitjançant moltes eines diferents i la majoria de cops quasi sense adonar-se, ofereixen punts de millora per a companyies i campanyes de màrqueting, permeten descobrir les necessitats dels usuaris i així millorar substancialment les decisions que l'empresa ha de prendre. També s'ha de tenir en compte que els beneficis per a les empreses són enormes, ja que per una banda faciliten l'avaluació dels productes o serveis que proporciona mitjançant l'anàlisi de dades, podent obtenir informació molt valuosa per crear altres productes nous o redissenyar-ne els que ja existeixen per a obtenir resultats més òptims. I per l'altra, la segmentació de clients permet personalitzar accions, de manera que les empreses poden orientar els seus serveis i satisfer les necessitats dels seus consumidors sent més específics. D'aquesta forma, aquestes millores faciliten l'accessibilitat i la fluïdesa de la informació dins de la mateixa organització, creant dinàmica de treball i alhora un increment d'eficàcia del treball.

Però encara hi ha alguns temes que cal perfeccionar per tal de no trobar impediments amb aquesta forma de treball. Perquè un dels principals problemes és la privacitat de les dades, ja que les dades poden ser emmagatzemades amb una finalitat i atorgades a una empresa o entitat en concret, aquesta, traient importància a la informació, pot creure's amb la llibertat de transferir-les a altres entitats gràcies a la facilitat que Big Data proporciona, i conseqüentment, aquestes dades deixarien de ser privades i estarien incomplint permisos i alhora la lleis. Però alguns d'aquests problemes, queden resolts quan les empreses són responsables de les finalitats amb que obtenen les dades i actuen en conseqüència, únicament. Encara que a aquest problema va lligada la incapacitació d'identificació de les dades, ja que en alguns casos la informació obtinguda no es relaciona amb el context i pot fer variar erròniament els resultats obtinguts, així com les deduccions finals a l'hora de treure conclusions. També s'ha de tenir en compte la incapacitat de tractament d'informació a temps real, encara que és l'objectiu final i que és el camí que s'està seguint per continuar l'evolució del Big Data, encara en moltes situacions no s'ha aconseguit, i això pot afectar en resultats o valoracions que calgui mantenir la informació perfectament actualitzada. Per poder aprofitar bé el que es proporciona, s'ha de centralitzar la rellevància d'informació, perquè no totes les dades són igual d'interessants per als diferents mercats i segons les situacions en que es valorin. Ja que no tenen la mateixa importància les dades obtingudes sobre un tema en concret en una part del món, que en una altra o per un sector concret que per un altre totalment oposat.

5.2 Explicació detallada de les diferents eines

En aquest apartat s'introdueix amb detall algunes de les eines Apache més rellevants en la situació de treball d'avui en dia, se'n fa una compressió del temari extensa i es valoren les opcions per poder entendre el perquè de les decisions preses a l'hora de defensar i posicionar una elecció.

L'elecció de que totes les eines sigui Apache és incentivada perquè presenta, entre d'altres característiques, ser programari de codi obert, lliure de privacitat i mantingut per una comunitat d'usuaris amb la supervisió d'Apache Software Foundation[3], quedant excent de remuneracions als autors. Per altra banda, al ser tota una comunitat qui desenvolupa el programari, és més popular i fàcil aconseguir ajuda o suport i extensible segons les necessitas.

Respecte les eines concretes, s'han escollit tres tipus d'Apache per fer-ne un aprofundiment, Apache Hadoop, Apache Spark i Apache Flink, ja que cadascuna aporta diferents característiques que s'esmenten a continuació.

Apache Hadoop és un framework de software que suporta aplicacions distribuïdes sota una llicència lliure, que permet el processament de grans volums de dades a través de clústers i usant un model simple de programació. A més, el disseny permet passar de pocs nodes a milers de nodes de forma àgil.

El sistema distribuït emprat utilitza una arquitectura Mestre-Esclaus (Master-Slaves) on un node central distribueix la les tasques pels altres nodes esclaus per realitzar la optimització dels recursos. Aquest sistema té una arquitectura principal on essencialment calen dues estructures.

Primerament hi ha el sistema d'emmagatzematge de fitxers que reparteix les dades entre cada node de la xarxa, que és conegut com a Hadoop Distributed File System (HDFS). On es redueix l'entrada i la sortida (E/S) a la xarxa, permetent l'escalabilitat i la tolerància a fallides. Respecte els elements importants del clúster del sistema d'emmagatzematge, per una banda hi ha el NameNode (sent el clúster Master) que regula l'accés als fitxer per part del client i controla el flux dels fitxer de cada Slave distribuint els recursos que arriben al clúster Master. Per altra banda hi ha el DataNode (sent els clústers Slaves) que són els responsables de llegir i escriure les peticions dels clients, i de realitzar les tasques encarregades pel Master.

[3] Apache Software Foundation (ASF) és una comunitat descentralitzada de desenvolupadors, que treballen cadascun en els seus propis projectes de codi obert. Els projectes Apache es caracteritzen per un model de desenvolupament basat en el consens i la col·laboració en una llicència de programari oberta i pragmàtica.

Per últim hi ha la implementació de l'algorisme MapReduce per fer els càlculs i el processament de la informació de forma distribuïda. És un procés batch, on no cal la interacció humana ja que permet d'una forma simple, dividir i paral·lelitzar el treball sobre grans volums de dades, abstraient la complexitat que hi ha en els sistemes distribuïts i on l'aplicació divideix en petits fragments de treball, cadascun dels quals es pot executar o tornar a executar en qualsevol node del clúster.

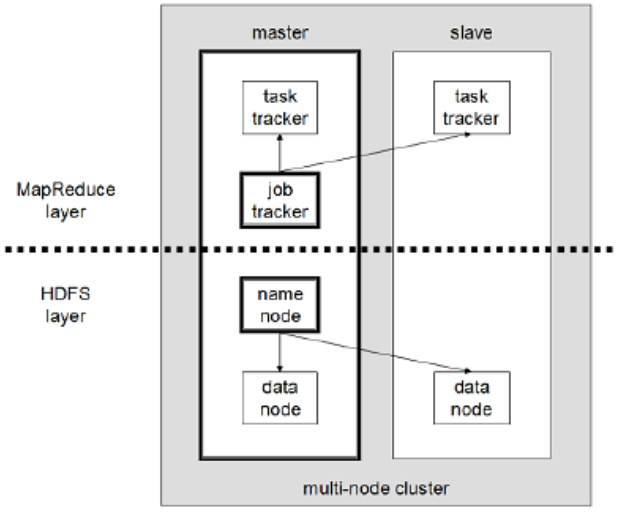


Fig. 2. Disseny de Hadoop amb un clúster de varis nodes.

Respecte els elements bàsics de l'algorisme, hi ha les funcions Map, que transforma un conjunt de dades claus-valors a una llista de nombres de parells, i on cadascun d'aquests elements es trobarà ordenat per la seva clau. I l'altra, és la funció Reduce, que s'utilitza per combinar els valors (amb la mateixa clau) en un mateix resultat.

Encara que, a la pràctica, quan es treballa amb un programa en MapReduce se sol conèixer també com a JobTracker (qui realitza la tasca de Master i treballa amb el clúster NameNode) i TaskTracker (qui realitza la tasca de Slaves i treballa amb els clústers DataNode), sent el punt d'interacció entre els usuari i el framework del MapReduce. Quan envien treballs MapReduce al JobTracker, els posa en una cua de treballs pendents i els executa en l'ordre d'arribada, gestionant l'assignació de tasques i delegant les tasques als TaskTrackers. Aquests executen tasques sota l'ordre del JobTracker i també manegen el moviment de dades entre la fase de Map i Reduce.

Apache Spark és un framework de processament ràpid i distribuït en memòria de segona generació que facilita la analítica de grans conjunts de dades integrant diferents paradigmes i en general, compatible amb Hadoop. Ja que pot funcionar en clústers de Hadoop o de manera independent, podent processar les dades en HDFS, HBase, Cassandra, Hive i qualsevol format Hadoop.

Spark manté l'escalabilitat lineal i la tolerància a fallides, perquè amplia les funcionalitats respecte MapReduce amb Directed Acyclic Grahp (DAG) i Resilient Distributed Dataset (RDD). Treballa dividint el sistema en varies capes, on cadascuna té una responsabilitat sent independents entre sí.

Per una banda, el DAG és un graf dirigit que no té cicles, és a dir, que per a cada node del graf no hi ha un camí directe que comenci i finalitzi en un mateix node, sinó que és un vèrtex que connecta a un altre però mai a si mateix i es va construint a mida que la consola d'Spark s'executa, ja que cada tasca crea un DAG d'etapes de treball en un determinat clúster. MapReduce crea un DAG amb dos estats predefinits (Map i Reduce) escrivint en disc els resultats de les etapes intermèdies entre Map i Reduce, en canvi els grafs DAG creats per Spark poden tenir qualsevol nombre d'etapes sent més ràpid, pel simple fet que no ha d'escriure en disc els resultats obtinguts en cada etapa intermèdia del graf.

L'altra funcionalitat d'Spark és RDD, que sorgeix quan les eines existents tenen problemes, cosa que produeix que es tractin les dades ineficientment a l'hora d'executar algorismes iteratius i processos de mineria de dades. En ambdós casos, mantenir les dades en memòria pot millorar el rendiment considerablement, però una vegada que les dades han estat llegides com a objectes RDD en Spark, poden realitzarse diverses operacions. Una de les quals es fer transformacions, que un cop aplicades, s'obté un nou i modificat RDD basat en l'original. O l'altra operació ,és fer accions, que consisteixen a aplicar una operació sobre un RDD i obtenir un valor com a resultat, que dependrà del tipus d'operació.

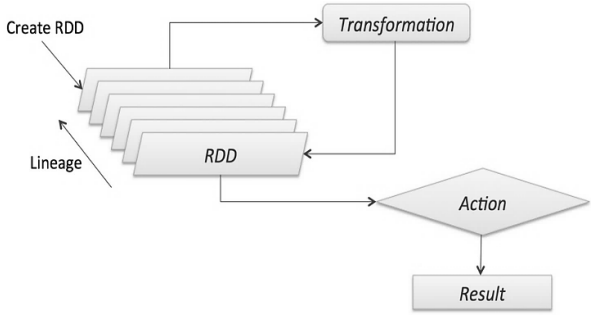


Fig. 3. Aplicació de la funcionalitat RDD d'Spark amb el fluxe de les operacions de transformació i acció.

Atenent que les tasques d'Spark poden necessitar realitzar diverses accions o transformacions sobre un conjunt de dades en particular, és recomanable emmagatzemar RDDs en memòria d'accés ràpid, com és el cas d'emprar la cache, on s'emmagatzemen les dades en memòria perquè no sigui necessari accedir en disc, agilitzant l'eficiència del procés.

Apache Flink és un framework de codi obert per a la analítica de grans dades distribuïdes, al igual que Hadoop i Spark. El nucli és un motor de flux de dades en streaming distribuït, amb l'objectiu d'optimitzar el MapReduce juntament amb el sistema de dades en paral·lel, ja que és compatible amb una amplia diversitat de dades, més enllà de si són parelles de claus-valors. Però Flink no proporciona un sistema d'emmagatzemament de dades propi, perquè les dades d'entrada han de ser emmagatzemades en un sistema distribuït com HDFS o HBase de Hadoop cosa que ens evita poder continuar comparant-lo com els altres dos tipus d'Apaches que actuen de forma individualitzada.

5.3 Requisits teòrics del Big Data per l'empresa

Quan una empresa es planteja la implementació de Big Data, s'ha de prendre el temps necessari per entendre els objectius del negoci. És vital establir entrevistes amb els nivells directius involucrats, ja que són els qui marquen la direcció dels esforços per a l'anàlisi de la informació. De no tenir aquest control, aplicar Big Data és un tasca que no produirà els beneficis esperats per una producció errònia.

L'empresa ha de qüestionar quin tipus de pregunta és la que es vol resoldre. Pot haver-hi preguntes tan bàsiques com "què va ser el que va passar", per tenir un punt de partida i per tenir un component del qual partir l'anàlisi.

La pròpia organització ha de ser crítica en l'avaluació dels recursos, ha de valorar el personal que es té, si és l'adequat per les tasques que es porten a terme, si la tecnologia és idònia o si hi ha els serveis pertinents per efectuar el treball. Si no es compleixen aquests recursos, l'empresa ha d'estar disposada a fer canvis tant en la forma d'actuar com en inversions per realitzar la implementació correctament.

Tenir una estratègia d'adquisició de dades, és a dir, com s'han d'adquirit les dades, si les infraestructures són aptes per donar resposta a les necessitats que hi ha i per poder realitzar el procés analític. A més de plantejar, si les fonts de les dades son fiables i confiables por poder realitzar l'estudi i les estadístiques. Quan es té el procés de recopilació de la informació, en paral·lel es treballa amb la part de la infraestructura que suportarà aquesta solució, per tal de tenir hardware que suporti les eines de treball.

S'ha de consultar i deixar clar amb totes les parts interessades, els objectius desitjats amb la creació d'un esquema de resultats anhelats, procediments automatitzats, informes. A més de tenir en compte la integració amb els controls de seguretat que s'hi veuran relacionats, així com la postulació detallada dels nivells de privacitat.

Un cop resoltes aquestes incògnites prèvies a la implantació, l'organització ha de procedir a l'elecció de l'eina correcta. Perquè la inversió sigui menor, s'ha d'escollir una eina de codi obert, per poder adaptar-la millor a les necessitats sense haver d'invertir en mòduls privats d'empreses que exigeixen exclusivitat en els productes. A partir de l'explicació detallada en els apartats anteriors de les diverses eines, les connotacions negatives filen tant prim que quasi són les mateixes per un sistema com per l'altre.

Apache Hadoop ha estat el primer framework de processament de gran volums de dades dissenyat per Google. Té molta flexibilitat i mutabilitat, ja que no es necessari conèixer i definir de forma exacta cóm són les dades abans d'incorporar-les al sistema. Permet accedir a la informació i processar-la independent del tipus, utilitzant qualsevol paradigmes i tecnologies, tant si son batch, iteratius, com si son MapReduce o nous sistemes com Spark, Mesos...

Apache Spark millora en quant a la computació en memòria com s'ha explicat als apartats previs, fent càlculs ens grafs i anant més enllà d'operacions en batch de MapReduce. Pot executar a temps real més ràpid que Hadoop pel que llavors, té una millor interactivitat i millor productivitat per analistes. Pot coexistir amb altres arquitectures Big Data, ja que al aparèixer quan ja hi havia un protagonista en el mercat com és Hadoop, ha hagut d'adaptar-se a tot el que pogués per no tancar-se portes i intentar tenir el màxim d'implementacions cosa que li està sent molt beneficiós.

Per tant, en aquest projecte s'aplica Hadoop, ja que gràcies a l'antiguitat del sistema aporta fiabilitat als usuaris. Va ser el primer framework que permetia optimitzar ràpidament moltes dades i és per aquest motiu que continuarem en aquesta línia i els resultats al llarg de tot aquest temps d'implantacions ha estat positiu.

Spark és justament el sistema que pot fer-li la competència, segueix la tendència de millorar l'algorisme MapReduce i sembla ser que s'han obtingut bons resultats de rendiment, però encara és molt nou com per poder competir amb la solidesa de Hadoop, que arrel dels anys i de la diversitat d'empreses i usuaris que l'utilitzen s'han anat aplicant millores i versions que l'han fet més competitiu.

Un altre tema que ens impedeix continuar pel camí d'Spark és l'escassa documentació que hi ha en comparació amb Hadoop. Amb el temps delimitat que hi ha per a realitzar aquest treball, si la informació no és tan accessible impedeix el progrés òptim de desenvolupament. Encara que no es una de les raons de pes, és un afegit per no aportar fiabilitat a l'hora de trobar suport en cas necessari, un cop s'estigui en la fase de construcció del software i es trobin casos concrets.

Per analitzar els resultats que processa Hadoop, que es complementarà amb l'Apache Hive, encara que hi ha multitud de complements de Hadoop que s'ajusten a les necessitats del moment com pot ser Hive, Pig o Spoop. En aquest cas, s'escull Hive per ser una infraestructura d'emmagatzematge de dades construïda sobre Hadoop que proporciona agrupació, consulta, processament dades, mitjançant un llenguatge de consultes semblant al SQL però que es denominat HiveQL, que s'orienta a realitzar DataWareHousing d'informació, gràcies als temps i als anàlisis que realitza amb les dades de l'empresa.

6 RESULTATS

6.1 Configuració i aixecament de Hadoop

Per realitzar l'anàlisis de dades, en aquest projecte s'utilitza una màquina sobre Linux, ja que és de codi obert, per tant es parteix d'un ordinador que té instal·lat Ubuntu 16.10, i es crea l'usuari Hadoop en cadascun dels nodes per treballar amb el clúster.

Ja en el terminal d'Ubuntu, s'ha d'instal·lar Java a l'última versió, per tal de procedir posteriorment a la configuració de Hadoop i Hive, a totes les màquines del clúster. Es realitza la següent configuració necessària:

A l'arxiu Hadoop-env.sh, es col·loca el PATH del JDK que es substitueix export JAVA_HOME=${JAVA_HOME} per export JAVA_HOME='/usr/lib/jvm/java-8-oracle'

A l'arxiu hdfs-site.xml s'afegeixen les rutes per la informació del Namenode i del Datanode, incloent:

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/
        namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/
        datanode</value>
</property>
```

A l'arxiu core-site.xml s'afegeix el nom del node Master.

```
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
```

S'ha de modificar l'arxiu yarn-site.xml, ja que permet sobreescriure un nombre de valors predeterminats per controlar els components del fil.

```
<property>
<name>yarn.nodemanager.aux-
        services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-
services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.Shuffle
        Handler</value>
</property>
```

A l'arxiu mapred-site.xml s'ha de sobreescriure un fragment per tal de controlar els components de l'execució dels treballs del MapReduce

```
<property>
<name>mapreduce.framework.name</name>
<value>yarn </value>
</property>
```

Seguidament, totes les màquines del clúster han de tenir accés automàtic entre sí i amb sí mateix a través del ssh amb l'usuari Hadoop, per això es genera la clau i s'afegeix a la llista de claus autoritzades, perquè Hadoop pugui accedir sense sol·licitar la clau cada cop.

```
ssh-keygen -t rsa -P " "
cat     $HOME/.ssh/id_rsa.pub     >>
$HOME/.ssh/authorized_keys
```

Un cop acabada la configuració, es pot començar a treballar amb Hadoop. S'ha d'estar situat a la carpeta con s'ha instal·lat Hadoop a la màquina Master i es realitza el format i configuració del Namenode (aquesta acció necessita internet per poder-se realitzar) i es creen els directoris namenode i datanode.

```
hdfs  namenode  -format
```

Finalment, s'aixequen els serveis del HDFS i MAPRED a la màquina master com es mostra a la imatge:

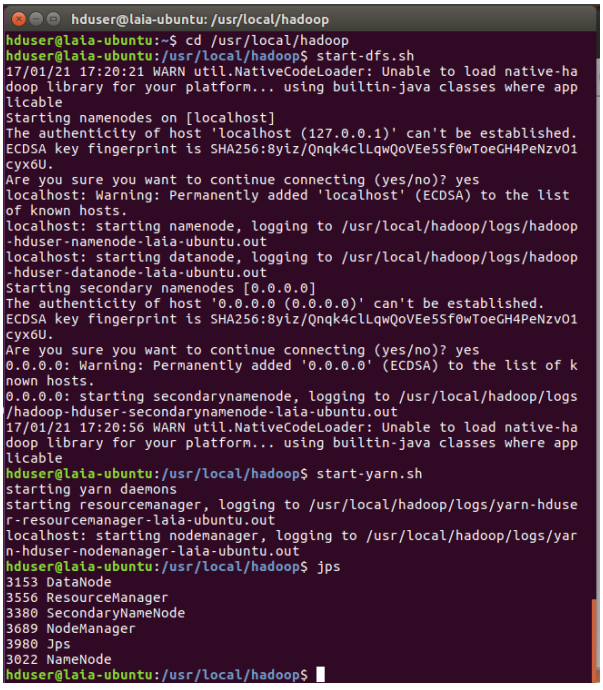


Fig. 4. Aixecament Hadoop.

Un cop feta la instal·lació i configuració, cada vegada que es vulgui iniciar Hadoop no caldrà fer tots els passos anterior, simplement introduir les comandes següents, s'accedeix al usuari hadoop, es cerca la carpeta on es troba i s'inicia.

```
sudo su hduser
cd /usr/local/hadoop
start-dfs.sh
start-yarn.sh
jps
```

6.2 Configuració i aixecament Hive

Amb Hadoop en marxa, ja que sense aquest Hive no pot treballar, es procedeix a la configuració de Hive amb les configuracions inicials.

```
cd apache-hive-0.13.0-bin/bin
export HIVE_HOME=$HOME/apache-hive-1.2.1-bin
export PATH=$PATH:$HIVE_HOME/bin
```

Es generen les carpetes HDFS per HIVE

```
cd apache-hive-0.13.0-bin/bin
export HIVE_HOME=$HOME/apache-hive-1.2.1-bin
export PATH=$PATH:$HIVE_HOME/bin
```

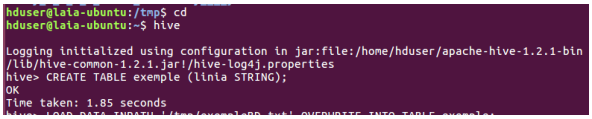
Una vegada tinguem el clúster corrent, Hive necessita dues carpetes dins del HDFS amb permisos de grup per poder funcionar, per crear aquestes carpetes i donar-los els permisos necessaris executarem les ordres següents:

```
hadoop fs -mkdir /user
hadoop fs -mkdir /user/hive
hadoop fs -mkdir /user/hive/warehouse
hadoop fs -mkdir /tmp
hadoop fs -chmod g+w /tmp
hadoop fs -chmod g+x /user/hive/warehouse
```

Amb l'arxiu TXT del qual s'han extret les mostres per tal de poder realitzar un exemple, amb un contingut de text d'unes 500 paraules aproximadament, que s'emmagatzema dins la carpeta bin, de la carpeta hive i es copia dins del HDFS.

```
hadoop fs -copyFromLocal exempleBD.txt
/tmp/exempleBD.txt
```

En aquest moment es quan s'executa la comanda hive i s'inicia l'execució.



Es crea una taula per emmagatzemar el fitxer de text, i seguit es carrega d'informació amb l'arxiu d'exemple. Cal tenir present que sempre a final de comanda amb Hive, s'ha d'indicar amb ";" sinó l'eina no ho contempla per execució.

```
hive> CREATE TABLE exemple (linia STRING);

hive> LOAD DATA INPATH
'/tmp/exempleBD.txt' OVERWRITE INTO TABLE
exemple;
```

Es genera una altra taula, per emmagatzemar els resultats del comptatge de paraules, on la funció de Hive "explode", pren una matriu com a entrada i dona sortida als elements de la matriu com a files separades, juntament amb el paràmetre "split" que consulta la ubicació exacta de la matriu.

```
hive> CREATE TABLE contador AS
SELECT paraula, count(1) AS compta FROM
(SELECT explode(split(linia,' ')) AS paraula
FROM text) w
GROUP BY paraula
ORDER BY compta;
```

Aquestes taules, generaran dos processos MapReduce per processar el treball, on el missatge resultant un cop finalitza el procés, és el següent i on es mostra el poc temps de resposta en analitzar del document de text.

```
[...]
Stage-2 map = 0%,  reduce = 0%
Stage-2 map = 100%,  reduce = 0%,
  Cumulative CPU 0.84 sec
Stage-2 map = 100%,  reduce = 100%,
  Cumulative CPU 2.12 sec
OK
Time taken: 42.304 seconds
```

Mostrant els resultats a la taula "contador", on apareix una columna amb el número de vegades que s'ha cercat la mateixa paraula en el document.

Per mostrar aquesta taula per pantalla es selecciona la taula.

```
hive> SELECT * FROM contador;
```

Aquest exemple senzill d'obtenció de paraules concretes amb la seva ordenació, sense tenir un document preparat per ser analitzat, sent simplement un text redactat, proposa demostrar que l'anàlisi de documents i dades en poc temps és factible amb les configuracions que s'han anat indicant al llarg del desenvolupament i del mostreig de resultats.

Amb l'elaboració de consultes més complexes, se'n poden extreure d'altres conclusions i decisions que poden ser vitals per prendre iniciatives a empreses o per l'analítica general de l'evolució de l'empresa.

7 CONCLUSIONS

En aquest apartat es pretén fer una valoració general després d'haver realitzat l'ampli estudi del projecte i un repàs de l'assoliment d'objectius que s'argumenten amb més detall durant la lectura d'aquest document i que consolida la satisfacció de poder ajudar als qui cerquin informació o guia en ell.

L'estudi mostra un seguit de plantejaments que quelcom es planteja quan es parla de Big Data i d'allò que l'envolta, així com la implantació i la documentació necessària per poder associar-ho a entitats o pymes, i que serveix en general per la gestió en l'avenç del tractat d'informació.

Un dels objectius inicials i necessaris per l'argumentar termes més específicament, és saber diferenciar i referir-se amb propietat a les diverses Bases de Dades, per poder escollir la millor opció per l'empresa o simplement per saber com catalogar allò que ja es té. Així com saber quin es el progrés evolutiu que segueix l'emmagatzematge de dades i en quin punt d'aquest s'està en cada moment i situació de l'empresa.

Seguidament, un altre objectiu principal que consta en saber valorar l'elecció de l'eina més idònia amb bases sòlides i argumentades, gràcies a la informació i comparació d'Apaches proporcionada en el estudi del projecte. Tanmateix com l'aprofundiment de Hadoop com a eina resolutiva en el pensament de software que permet el processament de dades. I d'igual manera, en com altres eines més noves segueixen l'exemple de Hadoop per continuar avançant i oferint línies alternatives però complementàries.

Com a objectiu, en un marc més teòric, es presenta la manera i els requisits previs que una empresa ha de tenir presents alhora d'implantar o començar a treballar amb eines de processament Big Data, adaptant l'analítica de les dades amb els objectius en que es centra, però sempre tenint-los present amb una finalitat concreta i sense divagar entre el munt de dades que es poden recollir al llarg del temps.

Finalment, la configuració i la petita demostració d'allò que s'ha estat esmentant durant tot el document demostra que sense grans inversions econòmiques, se'n poden treure cerques en pocs segons, cosa que si l'extensió de documents augmenten, el temps també ho farà però seguirà sent reduït en comparació al volum analitzat. A més, ofereix multitud d'utilitats en funció de les consultes i cerques que se'n facin de les dades.

7.1 Linies de millora

En aquest àmbit, el creixement i les modificacions són constants, per tant les millores i adaptacions poden arribar a ser extenses. A continuació s'esmenten algunes de les propostes interesants sorgides durant el desenvolupament.

La implantació de l'eina Hadoop pel processament Big Data en una empresa concreta, per adaptar requeriments addicionals que imposen amb uns anàlisis determinats i en l'àmbit concret que requereix cada empresa, i que per tant seria més exhaustiu.

L'elecció d'una eina de processament Big Data alternatiu a Hadoop, com podria ser Spark, per contrastar tot el que s'ha esmentat en un marc teòric. Inclús, analitzant les dades a temps real, per comparar els temps de resposta i si les modificacions de complements dels algoritmes funcionen.

8 AGRAÏMENTS

Sens dubte ha estat una etapa llarga, amb múltiples moments i situacions tan diferents, de les quals n'he après, tant acadèmicament com personalment, amb consells, reflexions i comentaris constructius dels qui s'han creuat amb mi durant aquests anys. La guia i la tutela d'en Ramón Musach Pi han estat pilars forts per poder realitzar aquest treball, gràcies als seus consells i sobretot al positivisme que ha transmès durant aquest temps, davant la incertesa plantejada en el transcurs del projecte.

Agrair en especial a Oriol Segura, per aquests anys de suport, per ser un punt clau en el meu dia a dia i per les seves aportacions amb preguntes incòmodes però que plantegen altres punts de vista interessants.

Nombrar, a Alba de la Fuente, qui ha estat la meva mà dreta durant aquest temps acadèmic i que naturalment és algú essencial en aquest moment.

Indubtablement, agrair els ànims, l'orgull i el suport incondicional dels meus pares, que han tingut la paciència i l'amor per fer-me arribar on sóc.

I per finalitzar aquests agraïments i aquesta etapa, retre gràcies a la gran família que tinc i als amics extraordinaris, que han vist l'esforç que he empleat en aconseguir un futur millor i, que plegats, gaudirem de la recompensa de completar aquesta etapa.

9 BIBLIOGRAFIA

[1] Steve Bryson, Exploring Gigabyte Data Sets in Real-time: Algorithms, Data Management and TimeCritical Design, Communications of the ACM, Volume 42, pp 82-90, Aug. 1999, USA. (Book style). Consultat: 9 de Novembre de 2016.

[2] «Fets del Big Data », 18 d'Abril de 2016. [En línia]. Disponible a: http://dataiq.com.ar/blog/16-hechos-de-big-data/ Consultat: 17 de Gener de 2017.

[3] «Metodologia Cascada», 25 de Setembre de 2016. [En línia]. Disponible a: https://es.wikipedia.org/wiki/Desarrollo_en_cascada. Consultat: 10 d'Octubre de 2016.

[4] «Centre tecnològic CTIC», 15 de Desembre de 2015. [En línia]. Disponible a: http://datos.fundacionctic.org/sandbox/ catalog/faceted/. Consultat: 21 d'Octubre de 2016.

[5] «Definició Big Data», 25 de Novembre de 2016. [En línia]. Disponible a: https://es.wikipedia.org/wiki/Big_data. Consultat: 21 d'Octubre de 2016.

[6] «Explicació de les bases de dades generals», 15 de Novembre de 2016. [En línia]. Disponible a: https://es.wikipedia.org/wiki/Base_de_datos. Consultat: 15 de Desembre de 2016.

[7] «Diferencies entre SQL vs NoSQL», 14 de Gener de 2014. [En línia]. Disponible a: http://www.thegeekstuff.com/2014/01/sql-vs-nosql-db/. Consultat: 15 de Gener de 2017.

[8] «Comparació de base de dades amb Big Data», 2015. [En línia]. Disponible a: https://www.asee.org/documents/zones/ zone3/2015/Comparisons-of-Relational-Databases-with-Big-Data-a-Teaching-Approach.pdf. Consultat: 5 de Gener de 2017.

[9] «Diferencies entre Big Data i base de dades», 26 de Juny de 2015. [En línia]. Disponible a: https://www.quora.com/What-is-the-difference-between-big-data-and-DBMS. Consultat: 20 de Novembre de 2017.

[10] «Possibles avantatges del Big Data», 15 de Novembre de 2015. [En línia]. Disponible a: http://www.silicon.es/las-cinco-ventajas-competitivas-que-aporta-el-big-data-49286. Consultat: 27 de Novembre de 2016.

[11] «Possibles problemes del Big Data», 14 de Novembre de 2012. [En línia]. Disponible a: http://rocreguant.com/posibles-problemas-del-big-data/351/. Consultat: 29 de Novembre de 2016.

[12] «Big data i com s'implementa a una empresa », 21 de Març de 2014. [En línia]. Disponible a: http://www.pymesy autonomos.com/inspiracionparatunegocio/que-es-big-data-y-como-se-implementa-en-una-empresa. Consultat: 15 de Desembre de 2016.

[13] «Tutorial de MapReduce i implementació Hadoop», 3 de Juny de 2015. [En línia]. Disponible a: https://www. tutorialspoint.com/map_reduce/implementation_in_hadoop.htm. Consultat: 7 de Gener de 2017.

[14] «Tutorial d'implementació de Hive», 17 de Març de 2016. [En línia]. Disponible a: https://www.tutorialspoint. com/es/hive/hive_installation.htm. Consultat: 10 de Gener de 2017.