Sistema d'Informació Big Data basat en Apache Kylin

Pau Pulido Sabidó

Resum—Actualment hi ha una gran quantitat de dades i cada cop disposem d'un temps menor per analitzar-la. Les dades són informació important per a la presa de decisions. Els sistemes d'informació basats en Business Intelligence ens proporcionen la solució als problemes que ens generen la gran quantitat de dades i l'absència de temps.
Aquest projecte és una solució a aquests problemes tècnics i a la necessitat de substituir les solucions actuals, que amb el pas del temps i l'augment de volum de dades han vist afectades el seu rendiment.
La solució proposada pren com a referència una solució implementada per a un client de l'empresa Pos Potential que actualment dona resposta als problemes plantejats, però cal donar una resposta al nou volum de dades.
El sistema disposa d'un clúster distribuït de Hadoop realitzat amb màquines Debian, d'una ETL, un cub de dades OLAP i d'un informe Excel per consultar les dades.

Paraules clau— Debian, Apache, Hadoop, HBase, Hive, Kylin, Excel, Pentaho Data Integration, Spoon, OLAP, Informe, Business Intelligence, Sistema d'Informació

Abstract—Nowadays there is a large amount of data and every time we have less time to analyze it. Data are important information for decision making. The information Systems based on Business Intelligence provide us with the solution to the problemes generated by the large amount of data and the absence of time.
This project is a solution to these technical problems and to the need to replace the corrent solutions, which over time and the increase in volume of data have been affected by their performance.
The solution proposed takes as a reference a solution implemented for a Pos Potential company client that currently responds to the problems proposed, but we must respond to the new volume of data.
The System has a distributed Hadoop clúster made with Debian Machines, an ETL, an OLAP cube and an Excel report to consult data.

Index Terms— Debian, Apache, Hadoop, HBase, Hive, Kylin, Excel, Pentaho Data Integration, Spoon, OLAP, Report, Business Intelligence, Information system

1 Introducció

POS POTENTIAL és una empresa que ofereix solucions úniques basades en "Big Data" per simplificar els processos de decisió i la planificació d'accions en gran consum, és a dir, processa aquest gran volum de dades, per a facilitar informes fàcils de comprendre. Pos Potential processa les dades creant scripts de carrega que s'executen manualment o automàticament, a diari, tot depenent del tipus de dades i de proveïdor d'aquestes dades. Un cop s'han carregat les dades a una base de dades, aquestes dades són carregades a un Cub OnLine Analytical Processing (OLAP) [1], el qual el processem per a poder accedir ràpidament a elles i així facilitar qualsevol tipus de consulta de l'empresa externa que vulgui accedir a les dades. L'empresa disposa d'àmplia experiència assistint als clients, on a partir d'aquesta experiència ha perfeccionat les eines necessàries per dur a terme projectes específics. L'empresa està constantment millorant els processos i cercant les tecnologies més recents per a una millor eficiència i simplicitat, el gran volum de dades de vendes de mercats com el dels Estats Units o la Xina i la necessitat de temps de resposta menors són els motius principals de l'inici d'aquest projecte que intentarà resoldre els problemes que hi ha a l'empresa de processament de fitxers i bases de dades amb un volum de dades gran.

Desenvoluparem un sistema d'informació distribuït utilitzant eines Open Source d'Apache sobre una instal·lació de Debian. S'utilitzaran tres màquines virtuals, una que serà el node màster i dues que seran nodes del sistema distribuït.

Generarem l'estructura d'un sistema d'informació distribuït utilitzant les eines d'Apache i configurant-les de tal forma que ens permetin tenir un model distribuït i implementarem un cub OLAP, creant el disseny, mesures, particions i agregacions pertinents.

Per realitzar-ho utilitzarem un fitxer històric on hi ha totes les dades de Sell-Out (vendes) d'un client de l'empresa. Per carregar aquest fitxer històric crearem de zero un procés Extract – Transform – Load (ETL) [2] amb Pentaho Data Integration i explotarem aquestes dades amb un informe Excel que crearem on podrem accedir a aquestes dades i generar consultes.

Els objectius són generar l'estructura d'un sistema d'informació distribuït i implementar un cub OLAP.

- *E-mail de contacte: paupulido19@gmail.com*
- *Menció realitzada: Enginyeria del Software*
- *Treball tutoritzat per: Oriol Ramos Terrades (departament de Ciències de la Computació)*
- *Curs 2017/18*

2 CONCEPTES BUSINESS INTELLIGENCE

Un cub OLAP és una base de dades multidimensional que permet mostrar als clients dades de vendes de les sortides de caixa registradora (Sell-Out) amb diferents referències, com el Distribuïdor, la Botiga, el Producte o el Temps.

Una taula de fets és la taula central d'un model en estrella d'un cub OLAP i que té un nivel de detall de l'informació que conté.
La taula de fets que s'ha realitzat conté detall de dades a nivel de producte, botiga, distribuidor i data.
La taula de fets conté les claus primàries de les dimensions que defineixen el seu nivel de detall i els indicadors, les unitats, el valor i el volum.

Una dimensió és el conjunt de dades que ens permeten filtrar o agrupar la informació que conté la taula de fets.
Per exemple la dimensió de temps es pot agrupar en diferents nivells de granularitat:

Fig. 1 Granularitat de la dimensió Temps

Un procés ETL és un procés que conté tres fases diferents; extracció, transformació i càrrega de dades.
La primera fase del procés és l'extracció de dades amb l'objectiu d'extreure les dades de l'origen.
La segona fase del procés és la transformació de dades amb l'objectiu d'aplicar funcions o modificacions a les dades de l'origen.
La tercera fase del procés és la càrrega de dades a les taules de les bases de dades un cop aplicades les diferents transformacions.

3 METODOLOGIA

Per realitzar el projecte s'han utilitzat dues metodologies diferents, una per la primera part del projecte, durant la implementació de la infraestructura del sistema d'informació, i una altra diferent per la segona, durant la implementació del cub OLAP.
Per realitzar la primera part del projecte que tractava d'implementar la infraestructura que s'ha utilitzat posteriorment per implementar el cub OLAP s'ha utilitzat una metodologia Scrum, que s'executa en cicles de curta du rada i de duració fixa, anomenats iteracions. En aquest cas la duració de cada iteració ha estat d'una setmana.
La metodologia Scrum és correcte per aquesta primera part del projecte ja que és molt flexible i permet estimar més fàcilment el temps per implementar una determinada funcionalitat [3].
S'han realitzat reunions de curta durada diàriament amb el meu tutor a l'empresa anomenades "Daily Meetings" on per una part, davant d'una pissarra de Kanban [4] s'ha explicat el treball realitzat des l'últim meeting, què es realitzarà fins el pròxim meeting i quins impediments hi ha o hi pot haver.
Una vegada comentats els impediments, si hi ha, s'ha fet un exercici de cerca de possibles solucions amb el tutor.
Per realitzar la segona part del projecte s'ha utilitzat la metodologia CRISP - DM (Cross Industry Standard Process for Data Mining).
Aquesta metodologia és actualment la que més predomina al realitzar projectes d'aquest tipus.
CRISP – DM [5] és un model de procés de mineria de dades que descriu els enfocaments comuns. Aquesta metodologia es divideix principalment en sis fases; Comprensió del negoci, comprensió de les dades, preparació de les dades, modelatge, avaluació i desplegament, tal i com es pot observar a la Figura 2.

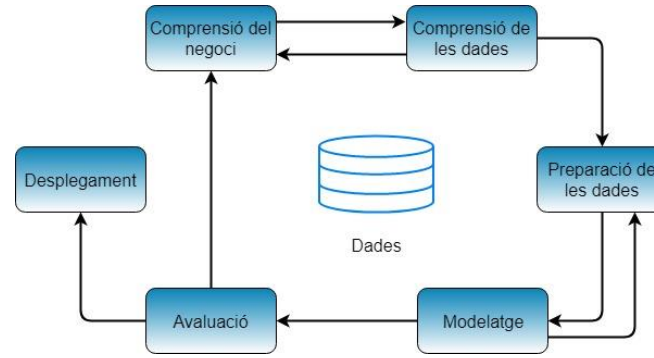

Fig. 2 Diagrama de la metodologia CRISP-DM

Com que s'ha utilitzat com a referència un sistema d'informació ja implementat, no ha calgut realitzar la primera fase de comprensió del negoci, per tant, s'ha començat la realització del projecte a la segona fase de comprensió de les dades.
Aquesta metodologia és adient pel nostre projecte degut a que té una forma iterativa que ens ha permés assolir tots els objectius.
Aquesta metodologia ens ha permés avançar i retrocedir si ha calgut entre algunes fases, fet que ha permés que si hi ha hagut un canvi ens els requisits o bé en els objectius ha permés tornar a la fase anterior sense haver de començar de nou el projecte.

3.1 Planificació

Per dur a terme el projecte dins dels terminis establerts s'ha dut a terme una planificació de les tasques i un diagrama de Gantt.
La planificació de les tasques i el diagrama de Gantt s'han realitzat amb el programa Microsoft Project [6] que permet l'administració de projectes.
S'ha hagut d'ajustar la planificació de tasques degut a imprevistos relacionats amb l'empresa. El tutor designat a l'empresa per realitzar aquest projecte va tutoritzar el treball només durant el primer mes degut a un canvi de feina, fet que va provocar un endarreriment a l'hora de realitzar les tasques.
Tal i com s'ha dit anteriorment, s'ha hagut d'ajustar la planificació i el Diagrama de Gantt que correspon a la

nova planificació està il·lustrat a la Figura 3.
Les tasques s'han dividit en quatre apartats corresponents al disseny i desenvolupament de la infraestructura, la ETL, el cub OLAP i l'informe.

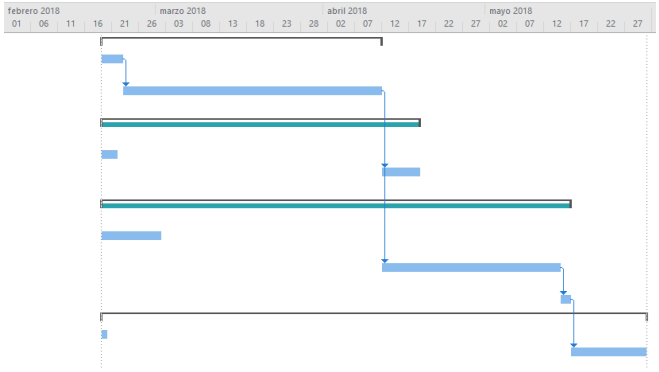

Fig. 3 Diagrama de Gantt

4 Estat de l'art

Actualment el departament IT genera informació automàticament, es rep un arxiu amb les dades i depenent del tipus de fitxer que sigui es carrega mitjançant scripts PHP o bé utilitzant SQL Server Integration Services (SSIS).
La informació d'aquests fitxers es carrega a la base de dades corresponent del client, una vegada carregat es processa el cub OLAP del client i ja es poden veure aquestes últimes dades que s'han carregat a SQL Server Analysis Services (SSAS) on hi ha implementat el cub del client.
Quan ja es poden veure les dades, el departament de consultoria és l'encarregat de generar els informes amb Excel connectant-se al cub OLAP o directament a la base de dades amb PowerPivot.

El departament IT també genera informes mitjançant SSRS (SQL Server Reporting Services).
Aquests informes s'allotgen a la pàgina web de l'empresa i PowerBI, on els clients es poden subscriure per rebre copies per correu electrònic periòdicament.

4.1 Sistema en funcionament

Actualment l'empresa utilitza un únic servidor on es duen a terme tots els processos relacionats amb l'emmagatzematge, la integració, l'anàlisi i l'explotació de les dades.

El servidor actual conté diferents tecnologies que realitzen aquests processos, totes les tecnologies són de Microsoft. El sistema operatiu que té el servidor en funcionament és Windows Server 2016 [7] i el sistema de gestió de base de dades relacionals és SQL Server, on hi ha allotjades les bases de dades..
La instancia de SQL Server instal·lada conté també SQL Server Integration Services (SSIS) [8] que és una plataforma per a la creació de solucions empresarials de transformacions i integracions de dades.

El servidor també conté SQL Server Analysis Services (SSAS) [9], on estan tots els cubs OLAP que té l'empresa implementats amb Visual Studio Data Tools [10]. SSAS és un motor de dades analítiques en linia que s'utilitza en solucions d'ajudo a la presa de decisions i Business Intelligence i proporciona les dades analítiques per informes empresarials i aplicacions client.

SQL Server Reporting Services (SSRS) [11] és l'eina que s'encarrega d'entregar les dades juntament amb Microsoft Excel [12] i PowerBI [13]. Mitjançant aquestes dues eines el client explota les dades que els hi proporciona l'eina i que l'eina extreu dels cubs OLAP de la instància de SSAS.

5 Situació proposada

El procediment actual però, té un cost alt de llicències i tenen limitacions quan la quantitat de dades a processar és d'una gran magnitud.
Per resoldre aquestes limitacions s'ha trobat una solució mitjançant eines Open Source que no tenen cap cost de llicències i que a més a més tenen un rendiment millor amb grans volums de dades.
Per a això, s'ha utlitzat Debian com a sistema operatiu on s'ha implementat les diferents eines d'Apache que s'han necessitat per realitzar el projecte: Hadoop, Hive, HBase i Kylin.
Apache Hive i Apache HBase són eines que s'han utilitzat per a l'emmagatzematge de dades i consultes.
A un nivell superior s'ha implementat Apache Kylin, que és el motor del cub que s'ha implementat.
Per a realitzar la càrrega de dades s'ha realitzat amb Pentaho Data Integration on s'ha implementat un procés d'ETL.
Aquesta solució té un millor rendiment degut a l'algoritme Map-Reduce que utilitza Hadoop.

El projecte Apache Hadoop desenvolupa programari de codi obert per a la computació confiable, escalable i distribuïda. Hadoop és un marc que permet el processament distribuït de grans conjunts de dades a través de clústers d'ordinadors mitjançant models de programació senzills. Està dissenyat per augmentar des de servidors individuals fins a milers de màquines, cadascun oferint còmput i emmagatzematge locals. En comptes de confiar en el maquinari per proporcionar alta disponibilitat, la biblioteca està dissenyada per detectar i controlar errors en la capa d'aplicació, de manera que ofereix un servei altament disponible en un clúster d'ordinadors, cadascun dels quals pot ser propens a fallades. [14]

Apache Hive és un projecte de programari de magatzem de dades integrat a Apache Hadoop per proporcionar resum, consulta i anàlisi de dades. Hive proporciona una interfície SQL per consultar data emmagatzemada en diverses bases de dades i sistemes de fitxers que s'integren amb Hadoop. Atès que les aplicacions de data warehousing funcionen amb llenguatges de consulta basats en SQL, Hive ajuda a la portabilitat d'aplicacions basades en SQL a Hadoop. [15]

HBase és una base de dades distribuïda no relacional de codi obert modelada a partir de Google Bigtable i escrita en Java. El seu desenvolupament forma part del projecte Hadoop de la fundació de programari Apache i s'executa sobre HDFS (el sistema d'arxius distribuïts de Hadoop). És a dir, proporciona una tolerància a falladles d'emmagatzemat de grans quantitats de dades disperses. [16]

Apache Kylin és un motor distribuït d'anàlisi de codi obert dissenyat per oferir una interfície SQL i una anàlisi multidimensional (OLAP) en Hadoop que admet conjunts de dades extremadament grans. [17]

Pentaho Data Integration és el component de Pentaho responsable dels processos ETL. [18]

6 INFRAESTRUCTURA

6.1 Arquitectura

S'ha implementat la infraestructura a un servidor local de l'empresa on hi ha instal·lat un ESXi Server, que és un software que ens permet crear vàries màquines virtuals i poder-les controlar des d'una interfície web.
L'objectiu és aconseguir que un fitxer de amb dades de ventes sigui tractat i es carreguin aquestes dades a les taules de les bases de dades corresponents.
El cub OLAP es connectarà i processarà les dades de les taules i quan el cub estigui processat es podran realitzar consultes, generar informes, etc.
Per a realitzar-ho, s'han creat tres màquines virtuals i s'ha instal·lat el sistema operatiu "Debian 9" a les tres, s'ha instal·lat la versió sense interfície gràfica, és a dir, funciona amb consola de comandes.
Per implementar la infraestructura s'ha seguit el disseny de la Figura 5 fet prèviament.
L'arquitectura conté una infraestructura d'un clúster de nodes de Hadoop, amb un node màster o principal i dos nodes fills.
S'han realitzat dos tipus diferents d'instal·lacions i configuracions a les màquines virtuals, depenent de si era un node màster o eren nodes fills.
Al node màster s'ha realitzat la instal·lació i configuració de Hadoop, Hive, HBase i Kylin.
Pels nodes fills s'ha realitzat només la instal·lació i configuració de Hadoop.
L'objectiu és separar on estan emmagatzemades les dades d'on es processen. Les dades estan emmagatzemades al node màster i els nodes fills són els que les processen.

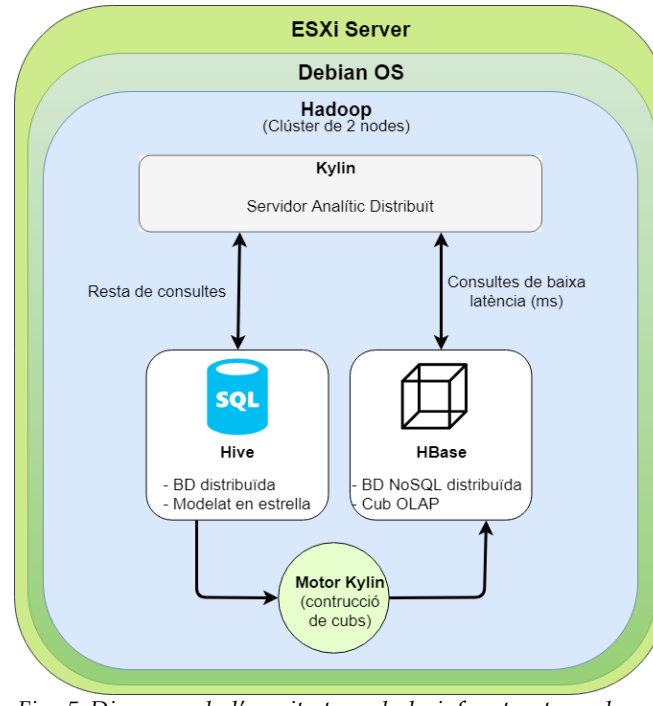

Fig. 5 Diagrama de l'arquitectura de la infraestructura al node màster

6.2 Creació ETL

S'ha implementat una ETL que donat un fitxer de vendes on hi ha el nom del producte venut, el seu EAN (número identificador del codi de barres), la marca del producte, el nom de la botiga, el país de la botiga, el nom del distribuïdor, la data, la quantitat i el preu.
Per a que el cub pugui retornar dades cal que les taules hi continguin dades.
Per fer això cal dissenyar i implementar una ETL, que és un procés d'Extracció, Transformació i Càrrega de dades (Extract - Transform - Load).
Prenent com a referència el fitxer de dades que es carrega mitjançant PHP a la base de dades que prendrem com a referència implementarem la següent ETL que hem dissenyat previament a la implementació tal i com es pot observar a la Figura 6.

El flux del procés s'inicia buscant el número EAN del producte a la taula de la dimensió producte, si aquest producte existeix, aquesta consulta retorna la clau primària de la taula anomenada "ProductKey", en cas contrari, s'insereix aquest producte com a una nova fila d'aquesta taula.
Aquesta etapa es realitza amb una consulta HiveQL, ja que no hi ha un paquet de "Lookup" o de cerca a una taula d'una base de dades.
Es realitza el mateix procediment amb la botiga, primer es busca si existeix i en cas afirmatiu la consulta en retorna la clau primària de la dimensió botiga, en aquest cas, anomenada "Storekey", en cas contrari, s'inserta aquesta botiga com a una nova fila de la taula creant una botiga nova.

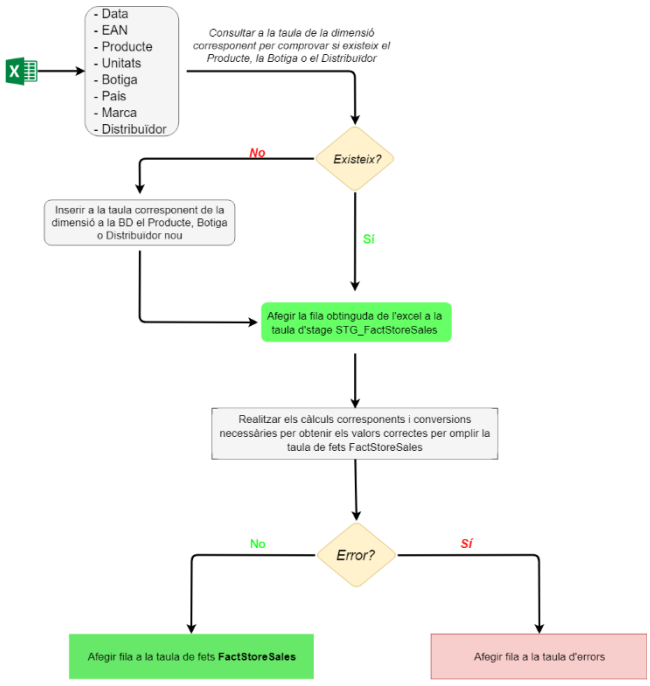

Fig. 6 Diagrama del procés d'ETL

Per últim es busca el distribuïdor a la taula de la dimensió distribuïdor per veure si existeix aquest, en cas afirmatiu, es retorna la clau primària de la taula anomenada "DataSourceKey", en cas contraris s'inserta una nova fila a la taula creant aquest distribuïdor.
Finalment, s'introdueix la venda d'un producte i una botiga determinats en un moment de temps determinat d'un distribuïdor com a una nova fila a la taula de fets "FactStoreSales".
La Figura 7 il·lustra la implementació de la ETL a Pentaho Data Integration amb els seus paquets corresponents.

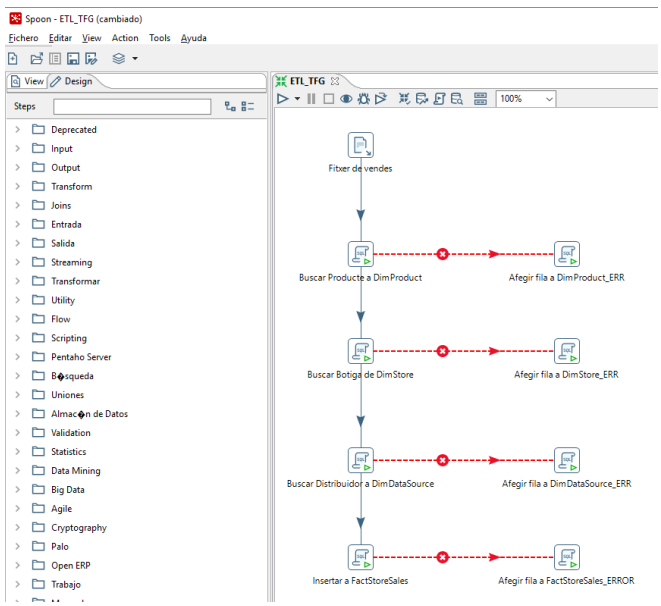

Fig. 7 Implementació del procés d'ETL a Pentaho Data Integration

Spoon no té paquets de Lookup o de cerca de registres a una taula d'una base de dades, per tant s'ha hagut de realitzar la cerca i l'afegiment de registres mitjançant consultes en llenguatge HiveQL a les diferents taules de dimensió i a la taula de fets, on també hi ha un control de duplicats per tal de no afegir més d'un cop una mateixa fila.

6.3 Creació Cub OLAP

6.3.1 Model del Cub

Per implementar el cub s'ha accedit a la interfície web de Kylin on es pot implementar un model del cub amb la seva eïna "Model Designer".
Primer de tot s'han carregat les taules de Hive a Kylin. Una vegada s'han carregat les taules, s'ha creat un model amb l'eina que ens proporciona Kylin, "Model Designer". En aquest model s'ha especificat el model de dades que tindrà, és a dir, les relacions que hi ha entre les taules de dimensions i la taula de fets. Aquestes relacions es fan mitjançant les claus primàries de les dimensions i les claus foranies a la taula de fets, com es pot observar a la Figura 8.

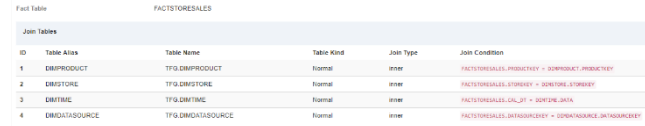

Fig. 8 Relacions de claus entre la taula de fets i les dimensions

S'han seleccionat les columnes de cada taula que es volen utilitzar com a columnes de dimensió i també els KPI's o mesures de la taula de fets que s'utilitzen com a indicadors, es poden observar a la Figura 9.

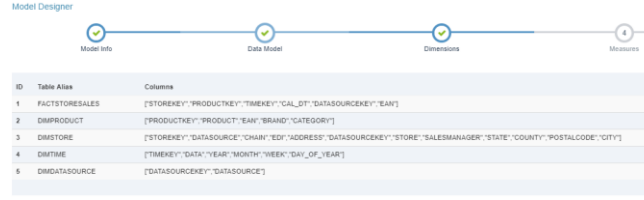

Fig. 9 Columnes de les taules que s'utilitzen com a filtres o mesures

A més a més s'han filtrat les files de la taula de fets tal que el valor de les vendes, el volum i les unitats venudes siguin més grans que 0. No tindrem en compte les devolucions, que tenen valors negatius, per tal de simplificar-ho.

6.3.2 Cub OLAP

S'ha implementat el cub a l'eina de Kylin que proporciona anomenada "Cube Designer".
Aquesta eina genera un codi JSON del cub que es pot utilitzar per implementar-ho a altres projectes de Kylin.
El cub té quatre dimensions diferents amb les seves claus primàries corresponents:

· Dimensió Producte - PK: ProductKey
· Dimensió Distribuïdor - PK: DataSourceKey
· Dimensió Botiga - PK: StoreKey
· Dimensió Temps - PK: PK_Date

El cub conté també una taula de fets anomenada FactStoreSales, es pot observar a la Figura 10.

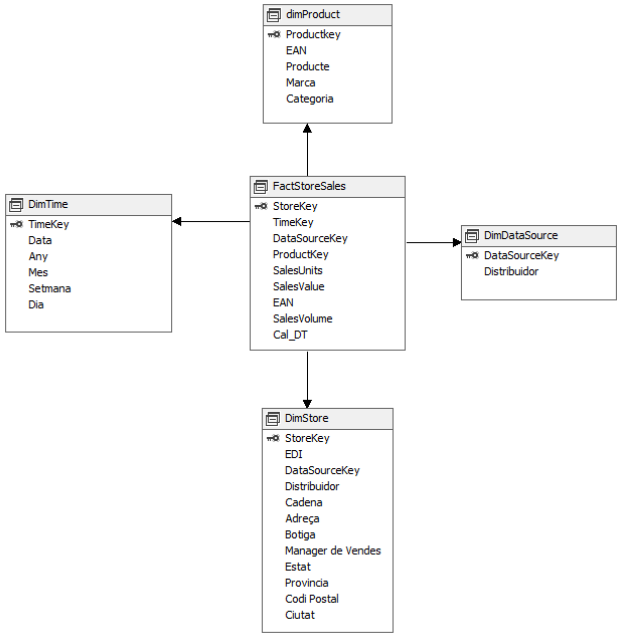

Fig. 10 Diagrama del model del cub

Primer s'ha especificat quin model s'utlitzarà per a la creació del cub, en aquest cas el model ha estat el model creat prèviament.

Tot seguit, l'eina ha demanat quins camps/columnes que contenen les taules de dimensions i la taula de fets del model creat s'han d'utilitzar al cub. També cal indicar quines mesures té el cub.

Aquestes mesures poden provenir de taules de dimensions, com pot ser en el nostre cas la mesura anomenada "Sales" que és el nombre de vendes que s'ha realitzat.
Les mesures també poden provenir de la taula de fets, com són les mesures "SalesValue", què és el valor de les vendes, "SalesUnits", que són les unitats venudes o "SalesVolume", què és el volum de les unitats venudes.

Finalment l'eina demana que especifiquem agregacions, que són utilitzades per millorar el rendiment del cub, i un tipus de motor que en el nostre cas és MapReduce de Hadoop.

Un cop s'ha realitzat tot el procediment a l'eina Cube Designer, s'haurà implementat el cub, tot i que, com no hi ha cap segment amb dades processades al cub encara no es poden realitzar consultes.
Per a que les consultes que es realitzin al cub ens retornin dades cal processar el cub especificant-li un període de temps.
Com que es vol tenir a l'abast totes les dades de vendes que s'han carregat, es processa el cub des l'inici de la primera dada fins la última.

6.4 Creació Informe

S'ha implementat un informe amb MS Excel on es poden observar diferents indicadors sobre les dades de la taula de fets, com el valor de les vendes o indicadors més el·laborats com els productes amb vendes.
Previament a la implementació de l'informe s'ha realitzat un disseny amb diferents filtres i indicadors, es pot observar a les Figures 11 i 12.
L'informe extreu la informació del cub creat a Kylin que es carrega al model de dades de PowerPivot.
Es pot observar a la Figura 13 com l'Informe ens retorna dades quan utilitzem els filtres de l'informe.

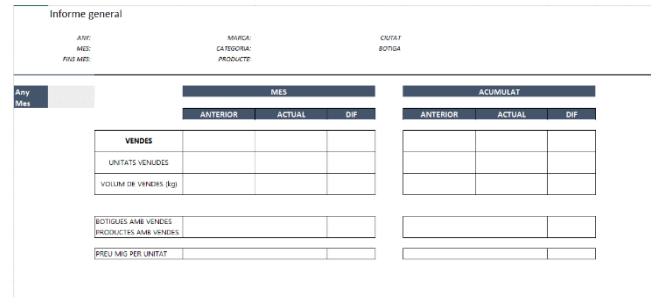

Fig. 11 Disseny de l'informe a Excel

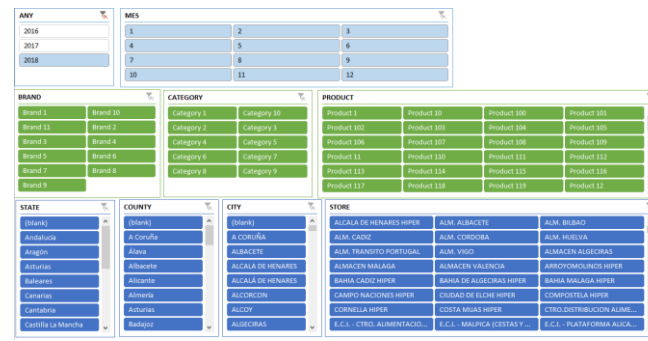

Fig. 12 Filtres de l'informe a Excel

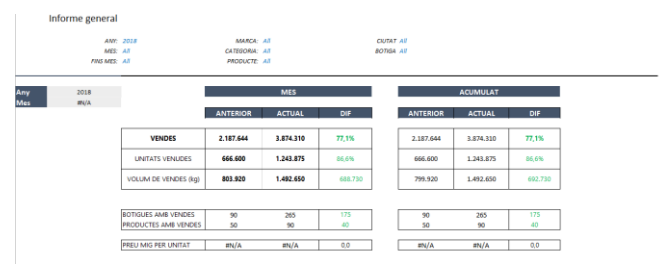

Fig. 13 Informe Excel

7 RESULTATS OBTINGUTS

Per presentar els resultats obtinguts utilitzarem dos criteris, per una part, els resultats obtinguts de la implementació del projecte i per una altra criteris de rendiment entre el cub OLAP creat a Kyin i el seu homònim amb SSAS, que crearem una copia de totes les dades i el cub i l'anomenarem CMS_DEMO per a que no hagi problemes de confidencialitat, les dades seran les mateixes als dos cubs.

7.1 Valoració dels resultats

S'han complert els dos objectius tècnics del projecte.
S'ha implementat tota la infraestructura de l'arquitectura dissenyada per poder realitzar solucions de Business Intelligence i s'ha implementat un cub OLAP a Apache Kylin com a solució.
Com a resultat extra, en un futur proper aquesta solució serà un recurs més per a l'empresa pel seu rendiment amb volums elevats de dades.

7.2 Proves de rendiment

Les proves de rendiment que s'han realitzat són tenint en compte el temps d'execució i resposta a les mateixes consultes als dos cubs que hi ha a l'empresa, al cub de Kylin i el cub de SSAS.
S'ha realitzat dos tipus de proves de rendiment amb els dos cubs amb el mateix número de registres.
El cub Kylin té un temps de processament i una mida d'espai en disc d'una gran magnitud en comparació al seu homònim, el cub SSAS.

La Figura 14 mostra amb detall la comparació entre el temps de processament i la mida per als mateixos registres i estructures molt similars del cub Kylin i el cub SSAS.

	Temps de processament	Mida del cub
Cub Kylin	21.28 minuts	744.68 MB
Cub SSAS	2.11 minuts	118.07 MB

Fig. 14 Comparativa temps de processament i de mida dels dos cubs

7.2.1 Taula dinàmica

S'ha realitzat una taula dinàmica a Excel amb cada cub OLAP, aplicant filtres i afegint totes les columnes de dimensions possibles i utilitzant totes les mesures creades.
El rendiment de les dues taules dinàmiques és similar, obtenen temps de resposta molt semblants.
Per tant, podem afirmar que no hi ha una millora d'un cub a un altre en aquest cas.

7.2.2 Consultes

S'han realitzat diverses consultes als dos cubs OLAP i s'ha utilitzat com a indicador de rendiment el temps d'execució i resposta de les consultes.
Com es pot observar a la figura 15, el temps d'execució de les consultes del cub SSAS és superior al cub Kylin, per tant el cub Kylin té un rendiment molt superior al cub SSAS.

	Consulta 1 (s)	Consulta 2 (s)	Consulta 3 (s)	Consulta 4 (s)
Cub Kylin	11.70	7.93	11.47	0.58
Cub SSAS	24.00	37.00	70.00	5.00
Diferència	-12.30	-29.07	-58.53	-4.42

Fig. 15 Taula de temps de diferents consultes

8 CONCLUSIONS

Les conclusions s'han extret en línia amb els objectius i els resultats obtinguts.
Com s'ha esmenat al principi, els objectius del treball de fi de grau eren generar l'estructura d'un sistema d'informació distribuït utilitzant les eines d'Apache i configurant-les de tal forma que ens permetin tenir un model distribuït i implementar un cub OLAP, creant el disseny d'aquest, les mesures, particions i agregacions pertinents.
Per una part, tenir una estructura distribuïda és complexa de realitzar i requereix un cost alt de temps però té un rendiment molt superior a una estructura no distribuïda.
Per una altra part, un cub OLAP a Kylin també té un cost alt de temps tant en la seva implementació com en el seu processament de les dades, però el seu rendiment és molt superior a un cub SSAS.
Respecte als resultats obtinguts, podem extreure una conclusió: el cub Kylin és més difícil de mantenir i té un cost de creació similar al cub SSAS, però, és molt superior en rendiment.

9 POSSIBLES EXTENSIONS

Com es tracta d'un rendiment tan superior, una extensió del projecte que durem a terme a l'empresa un cop s'hagi acabat el desenvolupament serà realitzar un cub a Kylin més complex amb més dimensions i particionant la taula de fets en vàries taules per realitzar proves de rendiment contra cubs de SSAS amb aquest disseny més complex.
Un altre element que no hem pogut tractar però que seria molt interessant com a extensió del treball és optimitzar el cub Kylin utilitzant "Cube Planner"[19] què és una eina que ens proporciona Kylin.
Finalment no hem tingut en compte que el hardware on hem realitzat totes les implementacions és inferior en rendiment al hardware on hi ha implementat el cub SSAS, per tant, també podríem suposar que en cas d'executar-se al mateix nivell de hardware que el cub SSAS el rendiment seria encara millor i la millora seria encara més gran.

AGRAÏMENTS

Amb la conclusió d'aquest article conclou també una etapa de la meva vida i m'agradaria agrair a totes aquelles persones que hi han format part d'una manera directa o indirecte.
Agraïr al meu tutor, Oriol Ramos Terrades, la seva ajuda i els seus consells.
Especialment m'agradaria agrair a la meva família, als meus amics i als companys de feina que m'han ajudat a realitzar a aquesta etapa de la meva vida i com a conseqüència, aquest treball.

BIBLIOGRAFIA

[1] Online Analytical Processing [Internet] . https://en.wikipedia.org/wiki/Online_analytical_processing [27 Jun. 2018].
[2] Extract Transform Load [Internet] . https://en.wikipedia.org/wiki/Extract,_transform,_load [27 Jun. 2018].
[3] An Empirical Framework For Learning (Not a Methodology). [Internet]. http://scrummethodology.com. [15 Mar. 2018].
[4] What is a Kanban Board? | LeanKit. [Internet]. https://leankit.com/learn/kanban/kanban-board. [15 Mar. 2018].
[5] Cross Industry Standard Process for Data Mining. [Internet]. https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. [27 Jun. 2018].
[6] Microsoft Project. [Internet]. https://es.wikipedia.org/wiki/Microsoft_Project. [27 Jun. 2018].
[7] Windows Server 2016. [Internet]. https://www.microsoft.com/es-xl/licensing/product-licensing/windows-server-2016.aspx. [27 Jun. 2018].
[8] SQL Server Integration Services. [Internet]. https://docs.microsoft.com/es-es/sql/integration-services/sql-server-integration-services?view=sql-server-2017. [27 Jun. 2018].
[9] About SQL Server Analysis Services. [Internet]. https://docs.microsoft.com/es-es/sql/analysis-services/analysis-services?view=sql-server-2017. [27 Jun. 2018].
[10] SQL Server Data Tools (SSDT). [Internet]. https://docs.microsoft.com/es-es/sql/ssdt/download-sql-server-data-tools-ssdt?view=sql-server-2017. [27 Jun. 2018].
[11] Reporting Services (SSRS). [Internet]. https://msdn.microsoft.com/es-es/library/ms159106(v=sql.120).aspx. [27 Jun. 2018].
[12] Microsoft Excel. [Internet]. https://es.wikipedia.org/wiki/Microsoft_Excel. [27 Jun. 2018].
[13] Power BI | Herramientas de BI para la visualización de datos interactivos. [Internet]. https://powerbi.microsoft.com/es-es/.
[14] Welcome to Apache™ Hadoop®! [Internet]. Hadoop.apache.org. [8 Mar. 2018].
[15] Welcome to Apache™ Hive [Internet]. Hive.apache.org. [8 Mar. 2018].
[16] Welcome to Apache™ HBase [Internet]. Hbase.apache.org. [8 Mar. 2018].
[17] Kylin A. Apache Kylin | Home [Internet]. Kylin.apache.org. [8 Mar. 2018].
[18] Pentaho Data Integration (Kettle) Tutorial – Pentaho Data Integration – Pentaho Wiki [Internet]. wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+(Kettle)+Tutorial. [8 Mar. 2018].
[19] Cube Planner [Internet]. http://kylin.apache.org/docs23/howto/howto_use_cube_planner.html. [22 May. 2018].
[20] Bringing interactive BI to big data Li Y. Bringing interactive BI to big data [Internet]. https://oreilly.com/ideas/bringing-interactive-bi-to-big-data [8 Mar. 2018]
[21] ANALYSIS BIG DATA OLAP SOBRE HADOOP CON APACHE KYLIN Analysis Big Data OLAP sobre Hadoop con Apache Kylin [Internet]. https://todobi.blogspot.com.es/2016/11/analysisbig-data-olap-sobe-hadoop-con.html [8 Mar. 2018]
[22] Online Analytical Processing on Hadoop using Apache Kylin Ijais.org [Internet]. https://ijais.org/archives/volume12/number2/ranawade-2017-ijais-4561682.pdf [8 Mar. 2018]
[23] KIMBALL, R., ROSS, M., THORNTHWAITE, W., MUNDY, J. Y BECKER, B. The Data Warehouse Lifecycle Toolkit Kimball R, Ross M, Thornthwaite W, Mundy J, Becker B. The Data Warehouse Lifecycle Toolkit. Hoboken: John Wiley & Sons; 2011.

APÈNDIX

A1. Interfícies eines Apache

A1.1 Hadoop Yarn

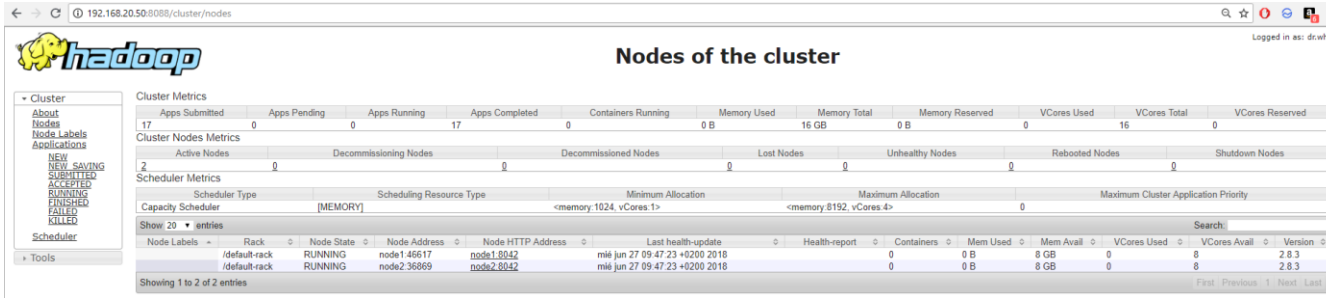

Fig. 16 Interfície Hadoop Yarn

A1.2 Hadoop DFS

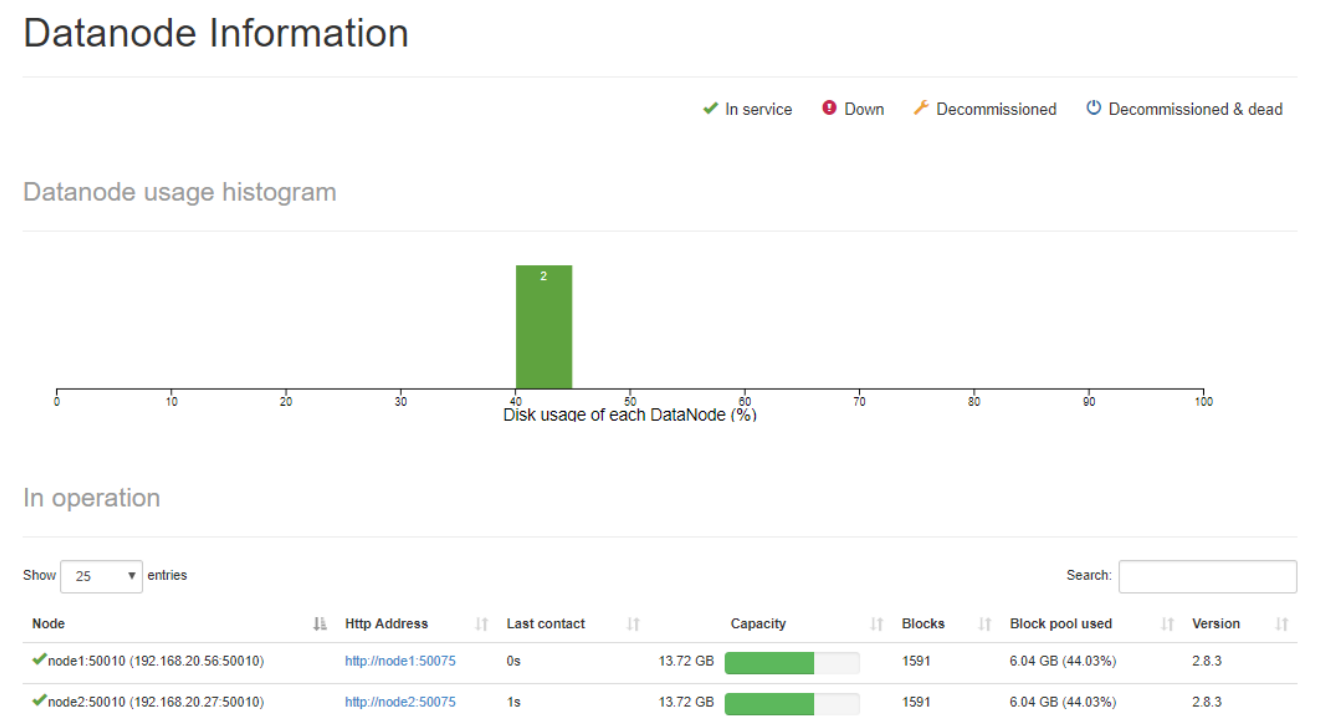

Fig. 17 Interfície Hadoop DFS

A1.3 HBase

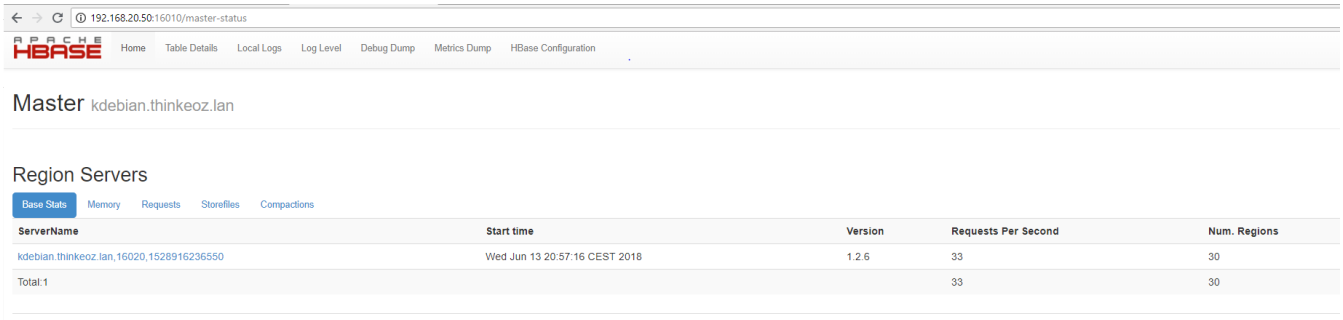

Fig. 18 Interfície HBase

A1.4 Hive

```
hadoop@kdebian:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-2.1.0-bin/lib/
SLF4J: Found binding in [jar:file:/home/hadoop/hadoop/share/hadoop/common
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an expla
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFac

Logging initialized using configuration in jar:file:/home/hadoop/apache-h
Hive-on-MR is deprecated in Hive 2 and may not be available in the future
hive> show tables;
OK
dimdatasource
dimproduct
dimstore
dimtime
factstoresales
kylin_account
kylin_cal_dt
kylin_category_groupings
kylin_country
kylin_intermediate_cub_tfg_61e3129e_def6_4f5f_a3bd_749e76bafb38
kylin_intermediate_cub_tfg_c0b8c164_fc96_4d34_933d_7788aa163963
kylin_intermediate_cube_tfg_b5e28e25_f801_4f9d_aca4_ef4877cbdae0
kylin_intermediate_cubotest_9d20c99f_6f81_4037_bc25_9e0912a419ce
kylin_intermediate_kylin_sales_cube_4a79f52f_62bf_46bf_a192_15d60648020b
kylin_intermediate_olap_cub_pau_71e43fcd_6538_4ec4_bc32_474b664452fe
kylin_intermediate_tfg_cubemodel_759d7ec4_115a_460c_9187_0d0ca2f48047
kylin_intermediate_tfg_cubemodel_b243b0c4_763f_4723_a508_0863843b0050
kylin_intermediate_tfg_cubemodel_f17b26af_ad6d_4aa8_ad81_52a152dbdd98
kylin_sales
Time taken: 0.63 seconds, Fetched: 19 row(s)
```

Fig. 19 Interfície Hive

A1.5 Kylin

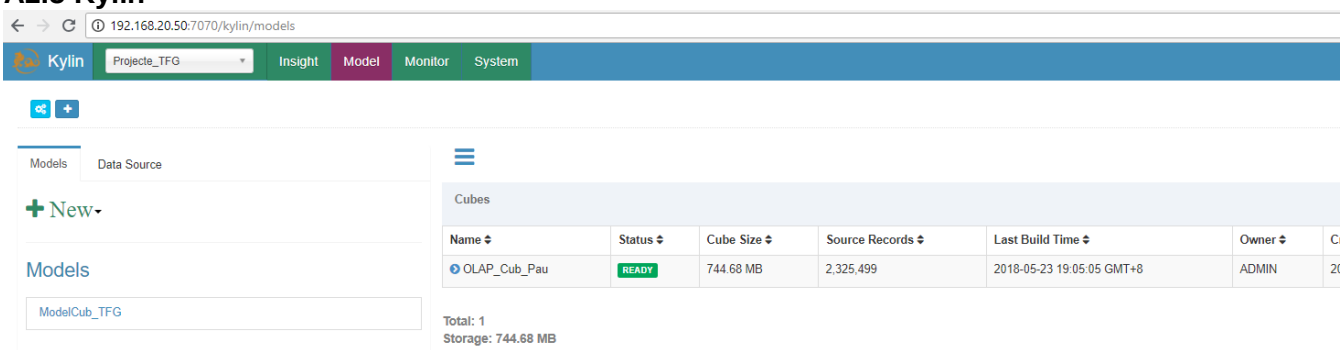

Fig. 20 Interfície Kylin

A2. CONSULTES ALS CUBS EN SQL

S'han realitzat les següents consultes SQL.

- Consulta 1:

```
SELECT * FROM FactStoreSales;
```

- Consulta 2:

```
SELECT * FROM FactStoreSales f inner join DimStore s on s.StoreKey=f.StoreKey;
```

· Consulta 3:

```
SELECT * FROM FactStoreSales f
Inner join DimStore s on s.StoreKey=f.Storekey
Inner join DimProduct p on p.ProductKey=f.ProductKey
Inner join DimTime t on t.data=f.cal_dt
Inner join DimDatasource d on d.datasourcekey=f.datasourcekey;
```

· Consulta 4:

```
SELECT year(cal_dt), month(cal_dt), f.storekey,f.productkey,sum(SalesValue) as SalesValue, sum(SalesUnits)
as SalesUnits, sum(SalesVolume) as SalesVolume
FROM FactStoreSales f
Inner join DimStore s on s.StoreKey=f.Storekey
Inner join DimProduct p on p.ProductKey=f.ProductKey
Inner join DimTime t on t.data=f.cal_dt
Inner join DimDatasource d on d.datasourcekey=f.datasourcekey
Where year(cal_dt)=2017
Group by year(cal_dt),month(cal_dt),f.storekey,f.productkey
Order by 1,2;
```

A2. CONSULTES HIVEQL

S'han realitzat les següents consultes HiveQL al procés ETL.

DimProduct:
"DECLARE @EAN nvarchar (13)=?
If (@EAN not in (select distinct EAN from DimProduct))
INSERT INTO DimProduct (ProductKey, Product, EAN) VALUES ((select max (Productkey) +1 from DimProduct), Product, EAN)"

DimStore:
"DECLARE @Store nvarchar (100)=?
If (@Store not in (select distinct Store from DimStore))
INSERT INTO DimStore (StoreKey, Store) VALUES ((select max (Storekey) +1 from DimStore), Store)"

DimDatasource:
"DECLARE @Datasource nvarchar (50)=?
If (@Datasource not in (select distinct Datasource from DimDatasource))
INSERT INTO DimDatasource (DatasourceKey, Datasource) VALUES ((select max (Datasourcekey) +1 from DimDatasource), Datasource)"

FactStoreSales:
"If ((Select count (*) from FactStoreSales where StoreKey = @StoreKey and ProductKey = @ ProductKey and TimeKey = @TimeKey and SalesUnits = @SalesUnits and SalesValue = @SalesValue and DataSourceKey = @DataSourceKey and EAN = @EAN) < 1)
INSERT INTO FactStoreSales (StoreKey, ProductKey, TimeKey, Cal_dt, SalesUnits, SalesValue, DataSourceKey, EAN)
VALUES (StoreKey, ProductKey, TimeKey, cal_dt, SalesUnits, SalesValue, DataSourceKey, EAN)"

A3. FILTRES A L'INFORME

S'han afegit filtres de tres dimensions diferents a l'informe.
De la dimensió Temps s'ha afegit filtre per any i per mesos.
De la dimensió Producte s'ha afegit filtre per Marca, Categoria i Producte.
De la dimensió Botiga s'ha afegit filtre per Comunitat Autònoma, Província, Ciutat i Botiga.

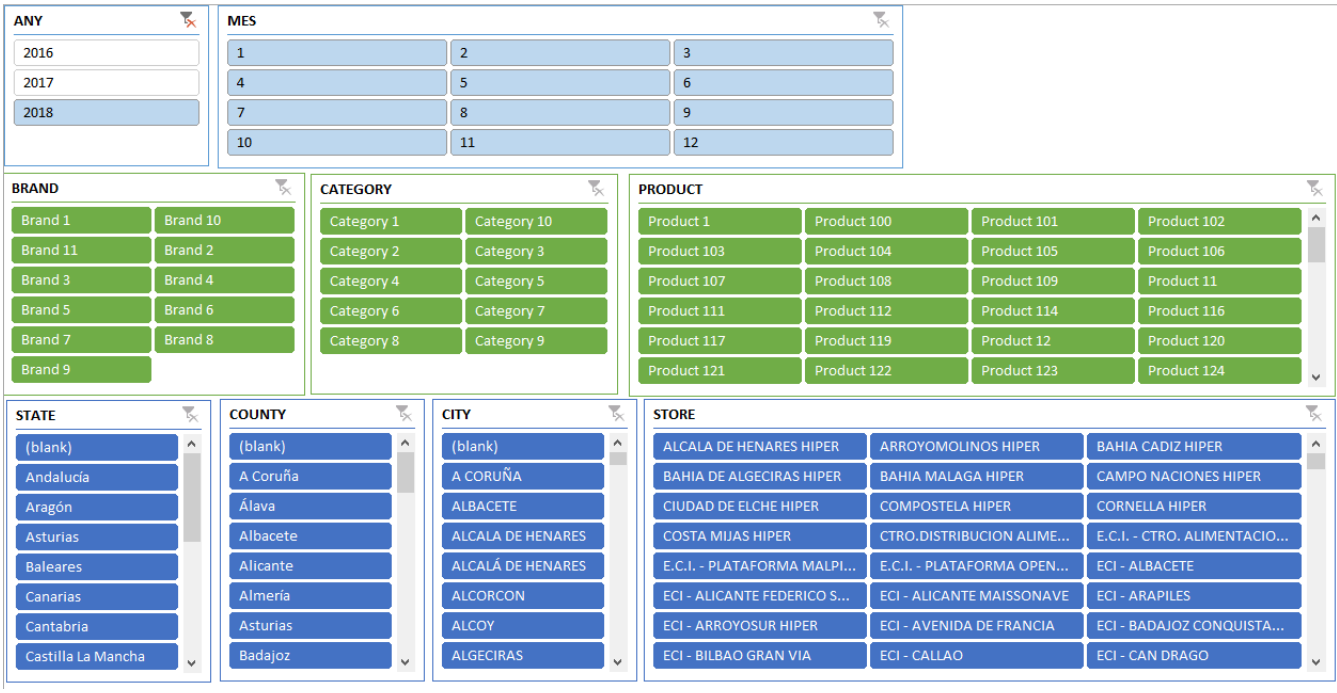

Fig. 21 Filtres de l'Informe Excel

A4. Taules dinàmiques

Taula dinàmica productes:

Row Labels	UDS Año Ant	UDS	UDS Dif %	PM Año Ant	PM	Valor	Distribución	Distribución Dif %
Product 111	8.308	9.268	11,56 %	0,00	7,13	66.115	240	5,73 %
Product 69	15.653	14.305	-8,61 %	0,00	4,03	57.695	244	6,09 %
Product 80	15.809	15.085	-4,58 %	0,00	3,04	45.818	249	-2,35 %
Product 2	27.111	26.226	-3,26 %	0,00	1,29	33.827	250	2,88 %
Product 72	2.973	4.628	55,67 %	0,00	7,23	33.447	228	13,43 %
Product 79	13.275	13.440	1,24 %	0,00	2,10	28.283	253	2,85 %
Product 48	9.865	7.970	-19,21 %	0,00	3,50	27.873	128	-10,49 %
Product 19	9.242	9.194	-0,52 %	0,00	2,92	26.841	218	2,35 %
Product 84	5.509	5.553	0,80 %	0,00	4,34	24.104	230	5,02 %
Product 70	4.570	3.823	-16,35 %	0,00	6,05	23.124	96	-41,82 %
Product 88	9.131	8.872	-2,84 %	0,00	2,42	21.453	249	-1,58 %
Product 67	4.278	7.148	67,09 %	0,00	2,63	18.776	242	23,47 %
Product 90	2.565	2.700	5,26 %	0,00	6,32	17.053	88	2,33 %
Product 27	3.140	3.103	-1,18 %	0,00	5,48	16.993	188	3,87 %
Product 14	7.116	7.343	3,19 %	0,00	2,22	16.272	226	0,89 %
Product 68	3.955	4.623	16,89 %	0,00	3,35	15.471	75	-1,32 %
Product 98	4.483	4.321	-3,61 %	0,00	3,54	15.313	124	-3,88 %
Product 87	11.538	11.579	0,36 %	0,00	1,32	15.291	247	-0,40 %
Product 124	790	1.602	102,78 %	0,00	9,53	15.268	212	46,21 %
Product 122		7.191			2,12	15.248	232	

Fig. 22 Taula dinàmica de productes a l'Informe Excel