Sistema d'intel·ligència de negoci per a l'anàlisi dels tractaments de reducció del colesterol

Elías Cid López
Pla d'estudis: Màster Enginyeria informàtica
Àrea de treball final: M1.221 - TFM-Business Intelligence

Consultor/a
David Amorós Alcaraz
Professor/a responsable de l'assignatura
Ferran Prados Carrasco
Data Lliurament: 07/017/2019

FITXA DEL TREBALL FINAL

Títol del treball:	*Sistema d'intel·ligència de negoci per a l'anàlisi dels tractaments de reducció del colesterol*
Nom de l'autor:	*Elías Cid López*
Nom del consultor/a:	*David Amorós Alcaraz*
Nom del PRA:	*Ferran Prados Carrasco*
Data de lliurament (mm/aaaa):	*01/2019*
Titulació o programa:	*Màster enginyeria informàtica*
Àrea del Treball Final:	*TFM-Business Intelligence*
Idioma del treball:	*Català*
Paraules clau	*Colesterol, BI, recerca*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

El client ha sol·licitat la implementació d'un sistema BI per a poder analitzar les dades d'un estudi realitzat a diversos pacients per a saber quin son els tractaments més efectius.

El resultat es un producte que conté un sistema de processos ETL per carregar i normalitzar la informació en una base de dades DW, per poder mostrar-la des de diferents punts de vista amb les tècniques de cubs, una base de dades on guardar tota la informació extreta dels estudis amb pacients i uns llistats mínims per fer el anàlisis.

Els llistats han de complir uns mínims per respondre a les preguntes analítiques que formulen els investigadors.

El projecte ha tingut una durada de 3 mesos aproximats, inclouen les fases de recerca de la plataforma, formació, disseny de la solució, implementació, proves i documentació corresponent.

Com a resultat hem implementat la solució que respon a aquests dubtes analítics proposats.

Con a conclusió final cap destacar que considero la eina PENTAHO coma una solució completa per a totes les parts d'un sistema BI. No obstant la part de presentació la podríem haver implementat amb altres solucions amb millor experiència d'usuari.

Abstract (in English, 250 words or less):

The customer has requested to implement a new BI system in order to analyse an investigation data done to several patients to know which the more effective treatments are.

The result must be a product which have a process system ETL to load and normalize the information in a database (DW), to be able to show it from different point of view within cube techniques, a database to save all the information charged from studies of patients and the minimal reports to do the analysis.

The reports must comply a minimal to answer all the analytics questions that ask the researchers.

The project has lasted around 3 month that's include platform research phase, training, solution's design, implementation, testing and documentation.

As a result, we have implemented a solution that answer the analítics doubts proposed.

To conclude I must highlight that I consider PENTAHO's suit like a full solution to implement all parts of a BI system. However, presentation part we could been implemented with another better solution with better user experience.

Con a conclusió final cap destacar que considero la eina PENTAHO coma una solució completa per a totes les parts d'un sistema BI. No obstant la part de presentació la podríem haver implementat amb altres solucions amb millor experiència d'usuari.

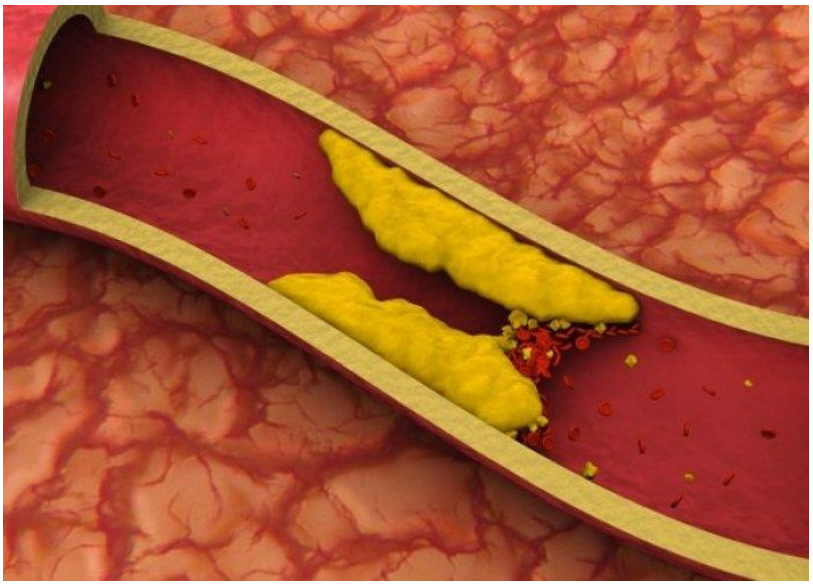
Índex

Llista de figures

1. Introducció

1.1 Context i justificació del Treball

El meu treball final de Màster està orientat en l'àrea de *Business Intelligence* a partir del qual construirem una eina de presa de decisió per analitzar diversos estudis en l'àmbit de la investigació de les malalties relacionades amb els nivells alts de colesterol (Hiperlipidèmia) [1]



Il·lustració 1 - Imatge obtinguda de la web https://sportadictos.com/2013/05/colesterol-alimentos-ayudan-controlarlo

El producte serà un entorn de BI on es pugui analitzar les dades de diversos estudis realitzats a pacients durant diferents tractaments contra el problema del colesterol alt.

El BI es coneix com un entorn on poder analitzar dades consolidades en un entorn anomenat data warehouse amb accés a les taules, vistes i principalment cubs multidimensional.[2]

1.2 Objectius del Treball

El client necessita una eina de Business Intelligence que l'ajudi a analitzar la informació dels pacients que han participat en l'estudi per investigar en les millores del tractaments de persones amb problemes de colesterol.

Els objectius principals son:

- Dissenyar i implementar un magatzem de dades i els processos ETL necessaris per adaptat la informació captada a les necessitats dels estudis.
- Implantar un sistema de BI per poder crear els reports corresponents.
- Implantar el nou sistema a la infraestructura del client per poder fer la carregar d'informació periòdicament
- Dissenyar i implementar els reports necessaris per poder respondre les qüestions marcades pel client.

Il·lustració 2 - Imatge obtinguda de la web
https://www.slpowers.com/BI-analytics

1.3 Enfocament i mètode seguit

Existeixen diferents estratègies a seguir per donar una bona solució i completament vàlida com podria ser implementar una aplicació a mida que abasti totes les parts del producte, des de la extracció i transformació de les dades fins a la presentació dels informes. Per mitjans de qualsevol llenguatge de programació es podria desenvolupar aquesta aplicació.

Però en els últims anys aquesta rama de la informàtica ha evolucionat molt i el mercat està ple de solucions fàcilment adaptables a quasi qualsevol necessitat de BI.

En el meu cas ens hem decantat per utilitzar diverses eines ja existents per desenvolupar la solució, de manera que la implementació tindrà una durada molt menor que no pas implementant de cero totes les parts del projecte.

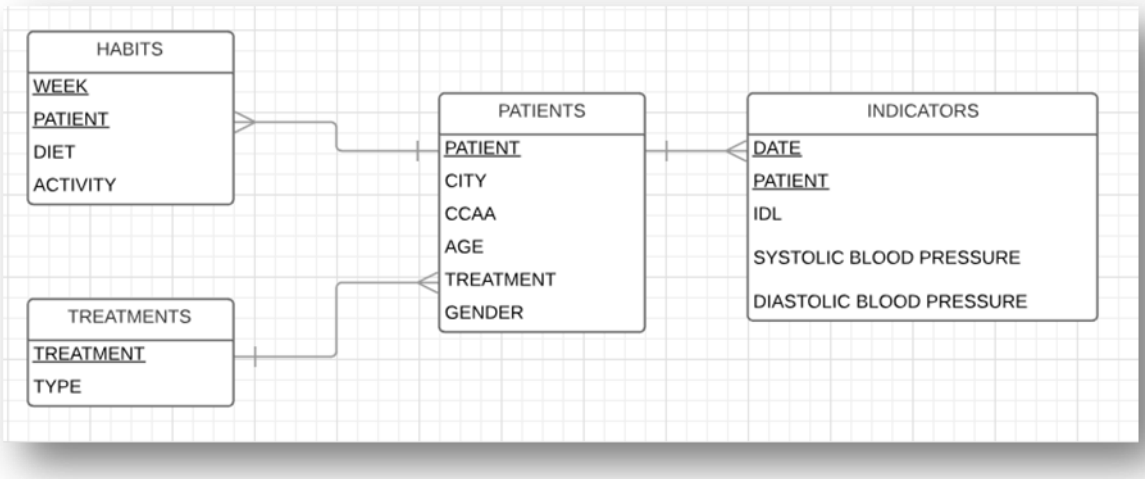
El client ha especificat els següents requeriments que s'han de complir per poder satisfer les seves necessitats relacionades amb els estudis de malalts de colesterol.

El sistema BI ha de poder respondre a les següents preguntes:

1. Quina és la relació entre els diferents tractaments i l'evolució dels pacients?
2. Existeixen teràpies més eficaces?
3. Ha influït en el resultat, els hàbits dels pacients?
4. L'evolució al llarg del temps, per un mateix tractament , depenen d'algun factor com els hàbits.
5. Hi ha diferències en el resultat d'un tractament segons el lloc geogràfic del pacient?
6. Hi ha algun període de l'any on el tractament sigui més o menys efectiu?

El sistema que s'ha d'implementar ha de ser una eina *open source*.

L'origen de les dades del sistema serà un fitxer *excel* amb 4 fulles amb la informació de:



Il·lustració 3 - Creació pròpia del diagrama ER de les entitats inicials

El sistema ha d'estar compost pels següents components:

a. Fitxers de carrega. A partir de la captura manual de les dades, el personal de la investigació omplirà els excels manualment i els deixarà a una carpeta d'un servidor.

b. Sistema de carrega. El programa de carrega anirà a buscar els fitxers mes actuals, farà la lectura i carregarà les dades a unes taules temporals al sistema.

c. Carrega de dades externes o mestres. El sistema recuperarà d'altres fonts les dades necessàries per completar-les.

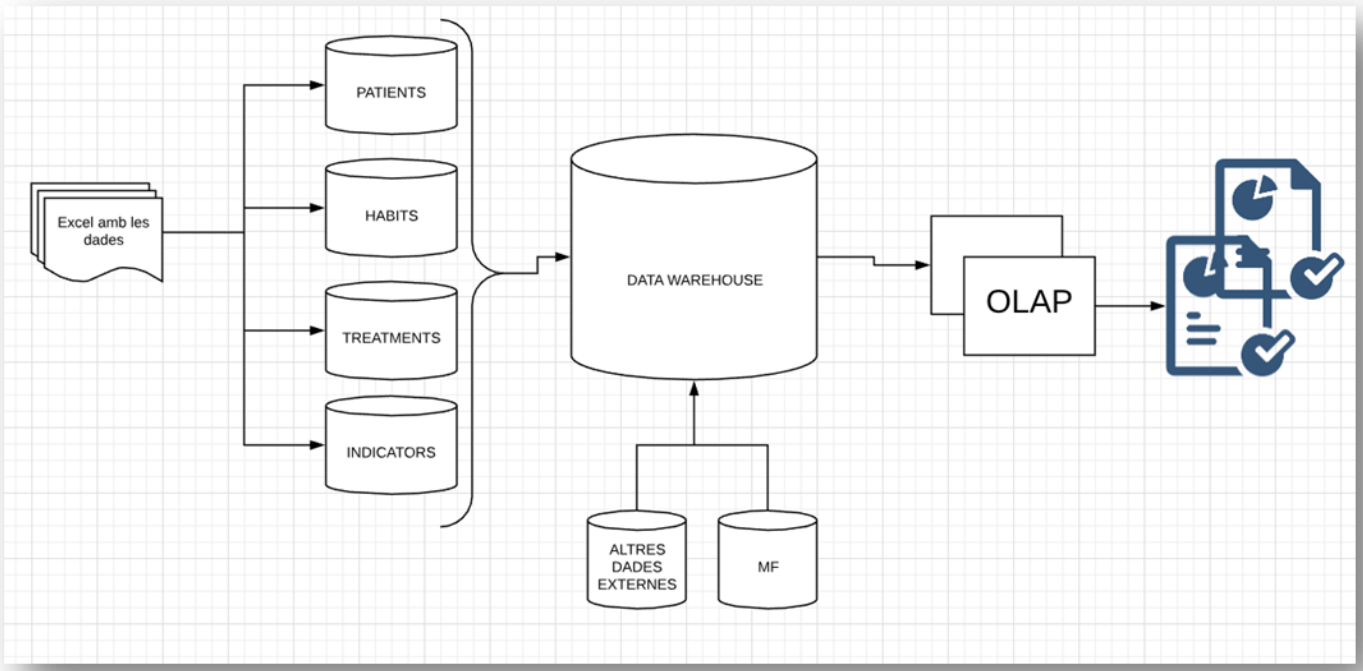
d. Classificació de les dades. El sistema farà les validacions necessàries per garantir la qualitat de les dades durant el seu anàlisis.

e. Transformació de les dades. El sistema farà les modificacions necessàries per millorar la presentació de la informació.

f. Depuració de les dades. El sistema ha de controlar que les dades obtingudes son consistents, revisant repeticions, inconsistències de tipus, etc...

g. Posteriorment l'usuari disposarà d'una eina per poder visualitzar les dades en la disposició necessària per poder fer el anàlisis.

Gràficament el sistema està representat de la següent forma:



Il·lustració 4 - Creació pròpia de l'enfoc global del sistema.

El disseny del reporting final no està definit, per tant només ha de complir que permeti respondre a les preguntes analítiques exposades anteriorment.

A continuació s'indiquen una llista de assumpcions que s'han de tenir en compte per poder portar a terme el projecte:

- El sistema està definit com una carregar "periòdica", es a dir que l'anàlisi es farà a partir de la captura de dades durant un període de temps i s'afegiran a les dades ja introduïdes en altres períodes d'estudis
- La generació dels fitxers es farà de forma manual pels investigadors. Ells mateixos recopilaran tota la informació la ompliran a l'Excel i deixaran el fitxer a una carpeta d'un servidor.
- El servidor on es guardi el fitxer haurà d'estar accessible des del sistema de carrega.

Per a realitzat la gestió i implementació del projecte utilitzaré el cicle de vida clàssic de gestió de aplicacions informàtiques basat en PMBOK. Les fases seran les següents:

- **Iniciació**. Aquesta fase en altres casos estaria basada en la preparació i selecció del tema del projecte. En aquest cas el consultor ha proposat el tema i s'ha facilitat a l'estudiant aquesta tasca.

- **Planificació**. De cara a garantir una bona execució del projecte es realitzarà mitjançant aquest document un pla de treball on es consideri la estimació de temps, costos i riscos de que es compon el projecte. Aquesta planificació serà orientativa i podria variar al llarg del temps durant l'execució revisant aquests canvis amb el consultor.

- **Execució**. En aquesta fase es treballarà en la construcció del producte que ha sol·licitat el client. Queden dins de l'abast el disseny i construcció del DW, la elecció de la eina BI, la construcció del scripts de ETL i la implementació dels reports com a solució final.
  Aquesta fase inclou la tasca de documentació a la memòria per no deixar-la pel final del projecte.

- **Seguiment i control**. Durant tota la execució del projecte es revisarà la planificació prèvia amb la realitat i es comunicarà amb el consultor per garantir que s'arriba a la finalització del projecte amb èxit. Un bon seguiment permet realimentar la informació que tenim del projecte amb tot allò que no havíem predit.

- **Tancament**. I per últim una fase molt important es poder tancar correctament el projecte amb una bona entrega al client, documentació i presentació del treball. La revisió de la planificació inicial amb el que ha estat la realitat ens ajuda a millorar per a futurs projectes.

[*] Informació estreta del document "TFM com a projecte" dels apunts de l'aula.

1.4 Planificació del Treball

Dedicació

Per fer una planificació el mes realista possible he calculat una dedicació setmanal de **20 hores**.

	HORES
L	2
M	2
X	2
J	2
V	0
S	4
D	8
HORES SETMANALS	20

A partir d'aquesta previsió, tenint en compte que totes les tasques del projecte les farà una sola persona sense possibilitat de paral·lelitzar, la dedicació total del treball des del proper 01/10/2018 fins el 07/01/2019 (data del lliurament final) serà de **280 hores**.

		HORES	Entrega	
W1	01-oct	20	PAC2	
W2	08-oct	20	PAC2	
W3	15-oct	20	PAC2	100
W4	22-oct	20	PAC2	
W5	29-oct	20	PAC2	
W6	05-nov	20		
W7	12-nov	20	PAC3	80
W8	19-nov	20		
W9	26-nov	20		
W10	03-dic	20	FINAL	
W11	10-dic	20	FINAL	
W12	17-dic	20	FINAL	100
W13	24-dic	20	FINAL	
W14	31-dic	20	FINAL	
W15	07-ene			
W16	14-ene			
W17	21-ene			
TOTAL		280		280

Taula de fites

Les tasques estan agrupades en les diferents entregues que té el TFM a partir de la durada estimada que té cadascuna.

Segons la dedicació prevista he estimat les hores que es podria dedicar a cada PAC segons les dates d'entrega:

#	Entregues	Durada / hores	Inici	Final	DUE DATE	Situació
1	PAC1 - Planificació	0	01/10/2018	01/10/2018	01/10/2018	DONE
2	PAC2	100	02/10/2018	05/11/2018	05/11/2018	PENDENT
3	PAC3	80	06/11/2018	03/12/2018	03/12/2018	PENDENT
4	Lliurament final	100	04/12/2018	07/01/2019	07/01/2019	PENDENT
5	Tribunal d'avaluació	8	14/01/2019	17/01/2019	17/01/2019	PENDENT

Amb aquesta dedicació he separat les següents fites en cada entrega segons la dedicació estimada:

#	Grup/Entrega Hores previstes	Fites	Durada / hores
1	PAC2 (100 hores)	Anàlisis dels requisits	6
2	PAC2 (100 hores)	Anàlisis i disseny del projecte	6
3	PAC2 (100 hores)	Disseny de les proves	14
4	PAC2 (100 hores)	Disseny del datawarehouse	14
5	PAC2 (100 hores)	Creació de la BBDD del DW	16
6	PAC2 (100 hores)	Implementació dels processos ETL	32
7	PAC3 (80 hores)	Anàlisis de les plataformes	24
8	PAC3 (80 hores)	I+D formació de la plataforma escollida	24
9	PAC3 (80 hores)	Implantació de la plataforma de BI	24
10	FINAL (100 hores)	Implementació dels reports finals	40
11	FINAL (100 hores)	Execució de la bateria de proves	8
12	FINAL (100 hores)	Lliurament del producte final	0
13	FINAL (100 hores)	Redacció de la memòria	40
14	FINAL (100 hores)	Lliurament de la memòria	0
15	FINAL (100 hores)	Preparació de la presentació	8
16		Defensa del TFM	
		Hores necessaries	256
		Hores disponibles	280
		Diferencia	-24

Es a dir, segons la planificació aquestes serien les tasques que s'entregarien a cada PAC:

<u>PAC2 – Lliurament el 05/11/2018</u>

- Anàlisis dels requisits
- Anàlisis i disseny del projecte
- Disseny de les proves
- Disseny del data warehouse
- Creació de la BBDD del DW
- Implementació dels processos ETL

<u>PAC3 – Lliurament el 03/12/2018</u>

- Anàlisis de les plataformes
- I+D formació de la plataforma escollida
- Implantació de la plataforma de BI

Entrega Final – Lliurament el 07/01/2019

- Implementació dels reports finals
- Execució de la bateria de proves
- Lliurament del producte final
- Redacció de la memòria
- Lliurament de la memòria
- Preparació de la presentació

Des del punt de vista econòmic podríem calcular a partir de la estimació de les 280 hores més les 8 hores dedicades a la planificació amb una tarifa de 60€/hora un preu total de **17.280 €** (Sense impostos)

Diagrama de gantt del 20/09/2018

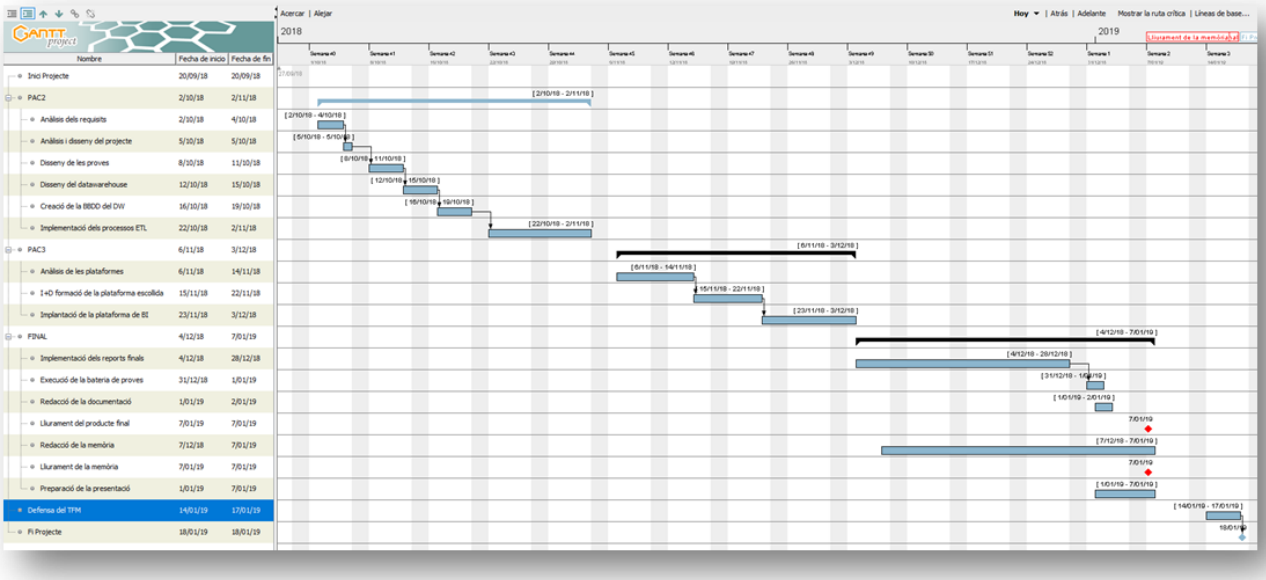


Diagrama de gantt del 5/11/2018 (Actualització)

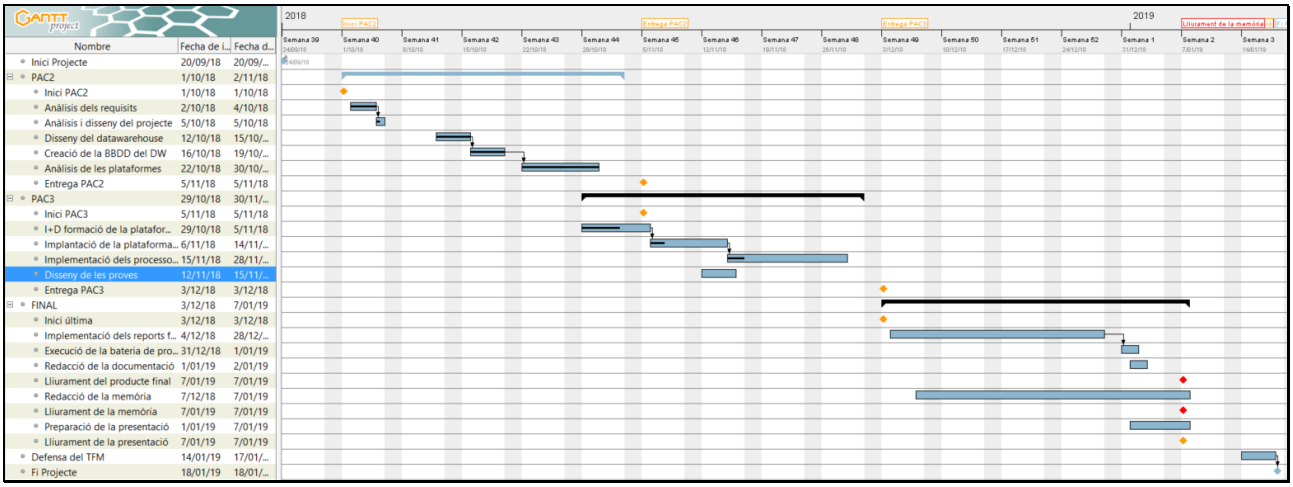
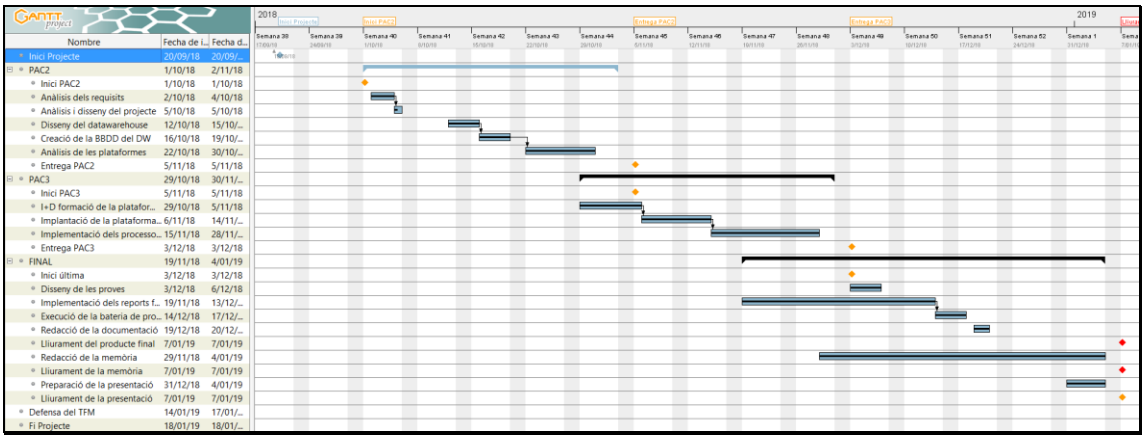


Diagrama de gantt del 04/01/2019 (Actualització)



El següent diagrama mostra la temporització de les tasques previstes. Document original (PLAN_ecidl_v1.gan) . Per poder obrir el fitxer es necessari la eina Open Source Ganttproject. Es pot descarregar des de la següent URL https://www.ganttproject.biz/

Riscos

S'han valorat els següents riscos a tenir en compte en la estimació del projecte:

R1. Eina BI.
Descripció: La eina encara s'ha escollit i es difícil predir la dificultat per a aprendre la seva utilització.
Mitigació: Es procurarà utilitzar alguna de les eines que sí hagi tingut experiència en altres projectes.

R2. Els festius de de Nadal.

Descripció: El període de la realització del projecte coincideix amb diverses dates festives de nadal amb diversos compromisos que podrien afectar a la planificació.
Mitigació: S'intentarà avançar alguna de les tasques per sobre de la dedicació setmanal prevista per evitar que afectin al calendari.

R3. Altres responsabilitats
Descripció: Al tenir dos fills i treballar a jornada complerta la dedicació al projecte està limitada a hores fora del horari laboral i les obligacions familiars. Aquesta dedicació podria no tenir un horari fixe per complir amb aquestes responsabilitats.
Mitigació: S'ha d'adaptar l'horari a aquestes responsabilitats.

1.5 Breu sumari de productes obtinguts

El sistema implementar està compost pels següents mòduls:

- **Entorn d'execució de les carregues**. El sistema *PENTAHO DATA INTEGRATION* permet automatitzar els processos implementats per a l'extracció, manipulació i carrega de les dades a analitzar.
- **Processos de carrega de les dades**. S'ha implementat un job i 3 transformacions per mitjans de la eina *Spoon* pròpia del entorn de disseny del *PENTAHO DATA INTEGRATION*.
- **BBDD Datawarehouse.** S'ha fet servir una BBDD MySQL per emmagatzemar les dades que es faran servir per a la presentació. La estructura creada té 4 taules temporals i 5 definitives per a la creació del cub.
- **Estructura del cub**. Per mitjans de la eina schema workbench s'ha implementat la estructura del cub necessària per a la posterior presentació.
- **Entorn de presentació**. El sistema que gestiona el cub i la presentació de les dades en llistats es el servidor *Pentaho business analytics*
- **Llistats**. S'ha implementat una sèrie de llistats a partir del cub creat per a poder respondre les preguntes necessàries per a complir els requeriments.

1.6 Breu descripció dels altres capítols de la memòria

- **Capítol 2.1 Disseny tècnic** . Documentació tècnica prèvia a la implementació del sistema BI
- **Capítol 2.2 Descripció del producte** . Descripció detallada del producte obtingut.

2. Resta de capítols

2.1. Disseny tècnic

2.1.1 Entorn tecnològic

Durant la fase inicial del projecte s'ha fet un estudio de les diverses eines que ofereix el mercat per aquest tipus de projecte.

Un dels requisits que sol·licita el client es que la eina suportada sigui open source, per tant es una de les característiques més importants en les que s'ha basat l'estudi.

Per fer la selecció de l'entorn hem dividit el sistema en 3 parts:

- La carrega i transformació de les dades
- L'emmagatzematge de la informació. DW
- La presentació de la informació.

Opcions de la part de Carrega i transformació de les dades.

- **Scripts de Unix**. Es pot arribar a gestionar la carregar de fitxers a BBDD per mitjans de comandes de unix o DOS (processos batch). Les instruccions per accedir a la base de dades dependrien del gestor DB escollit. [3]

- **Oracle data integrator**. L'eina d'Oracle serveix per definir, dissenyar i executar la carregar de dades en la bbdd. Aquesta eina es descarta per tenir una llicencia comercial. [4]

- **Qlik sense**. Una eina de la empresa qlik especialista en sistemes analístics i de BI. Actualment disposa d'eines per a gestionar la carregar de dades a partir de fitxers en el seu sistema. Un dels contres es que el sistema es bastant tancat en quant a la resta de parts pel que s'hauria de fer servir les mateixes eines per la capa de presentació. [5]

- **Microsoft flow**. Eina de Microsoft que es pot fer servir per automatitzar qualsevol acció que estigui dins del ventall de serveis que integra. Dona moltes facilitats a la integració i carrega de dades però un dels requisits es que les dades ja estiguin al núvol per poder treballar amb elles. [6]

- **Pentaho data Integration**. Eina open source que posseeix una suit integral per a totes las parts del projecte. Es la eina escollida per la gran comunitat que hi ha darrera i la seva facilitat per trobar documentació. [7]

Opcions de la part de emmagatzematge.

- Oracle 18c. El gestor de base de dades comercial mes famós actualment. Les seves llicencies son molt costoses per a un projecte com aquest. [8]

- MySQL 8.0x. El gestor de base de dades open source i de lliure us propietat d'Oracle corp. Considero la eina apropiada per fer-la servir de magatzem per a la informació del projecte. Per la facilitat d'un i de instal·lació es el gestor escollit per aquest projecte. [9]

- Vértica. Base de dades distribuïda pensada per a l'analisis de dades del BIG DATA. No es una bbdd relacional i s'adaptaria mijor si el volum de dades a carregar fossi molt mes alt. [10]

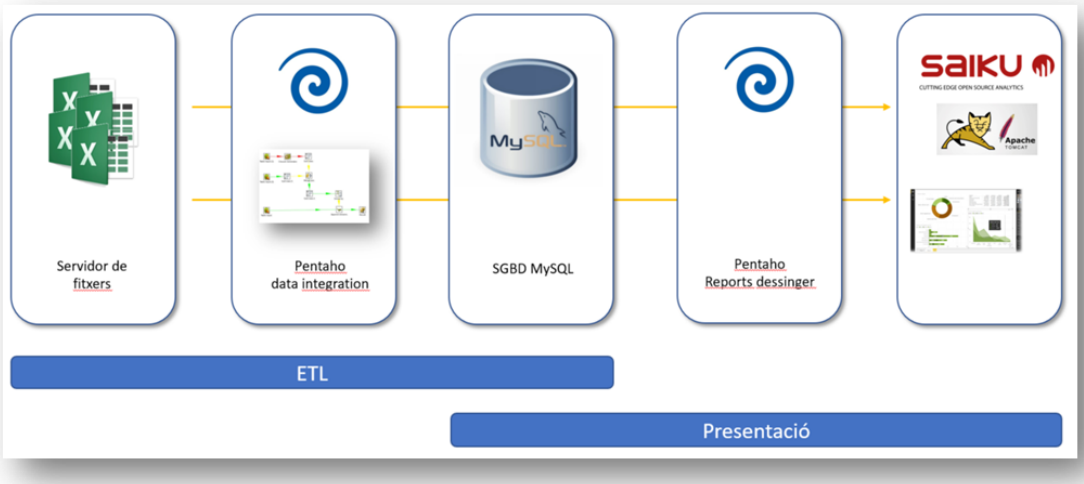
Opcions de la part de presentació.

- Microsoft powerbi. Entorn de BI que engloba totes les fases de projecte de BI, des de la carrega a la presentació final. En aquest cas es podria utilitzar solament la capa de presentació per la creació dels llistats. Disposa de solucions d'escriptori i al núvol. El sistema es distribueix a la part d'escriptori com a solució lliure però amb certes limitacions de compartició amb la organització la qual solament es pot fer servir des de la versió amb la llicencia de pagament corresponent [11]

- Pentaho. El siut de HITACHI també disposa de un mòdul especial per a la representació de les dades a partir de un DW o datamart preparat per a tal objeciu. Es l'escollit per a aquest projecte. [12]

- Qlikview o qlik sense. Eina de la empresa Qlik que engloba en la mateixa suite totes les fases del BI. Es una eina comercial encara que es pot utilitzar la personal Edition per obtenir les mateixes solucions amb limitacions de compartició. [13]

El següent gràfic representa des de un punt de vista global l'entorn tecnològic que es farà servir al projecte:

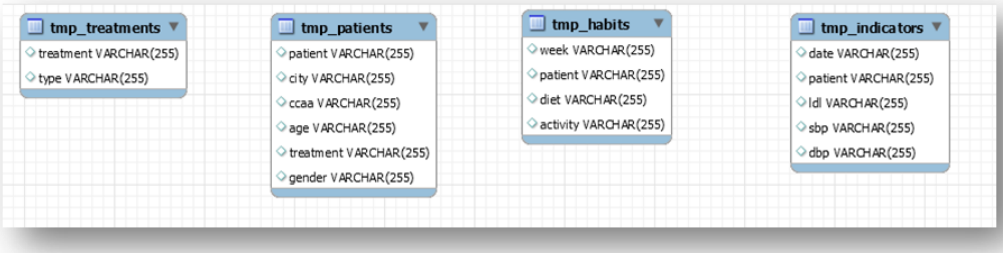


Il·lustració 5 - Creació pròpia de l'enfoc aproximat del sistema.

2.1.1 Disseny de BBDD

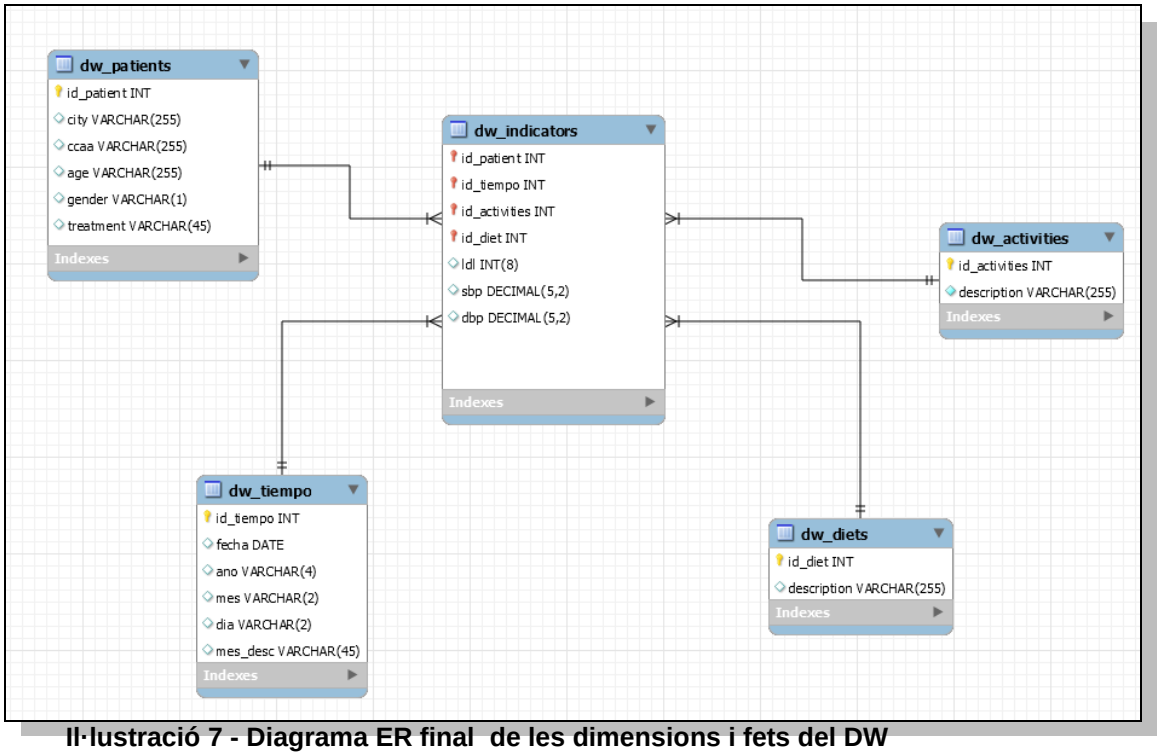
Després de fer el anàlisis de l'entorn tecnològic el gestor de BBDD escollit es MySQL. El sistema ha de tenir un servidor amb MYSQL 8.0x instal·lat i al qual es pugui accedir des de on estigui instal·lat el servidor de Kettel Pentaho.

La carrega de dades es farà directament en taules temporals des d'on es treballarà la seva depuració.



Il·lustració 6 - Diagrama ER de les taules temporals

Posteriorment s'actualitzaran als mestres corresponents.



Il·lustració 7 - Diagrama ER final de les dimensions i fets del DW

A partir d'aquest model de dades s'ha analitzat les necessitats dels cubs o agregades tenint en compte les següents dimensions i mesures per a cada qüestió que ha de resoldre el client:

Quina es la relació entre els diferents tractaments i l'evolució dels pacients
Mesures: Nivell de LDL
Dimensions: Tractament, mes, setmana (L'evolució es podrà observar millor si agreguen a nivell de setmana.)
S'ha de analitzar la mitjana del valor LDL per setmana, mes i tractament.

Existeixen teràpies mes eficàcies?
Mesures: Nivell de LDL, tenint en compte un valor de eficàcia com pot ser la variació en un mateix pacient.
Dimensions: Teràpia
S'ha de poder analitzar la mitjana de totes les variacions per a cada teràpia.

Ha influït en el resultat els hàbits dels pacients?
Mesures: Nivell de LDL, tenint en compte un valor de eficàcia com pot ser la variació en un mateix pacient i la diferencia contra el mateix valor per a cada hàbit.
Dimensions: Hàbit, teràpia.
S'ha de poder analitzar la diferencia entre els diferents hàbits i sense tenir el compte els hàbits.

L'evolució al llarg del temps per a un mateix tractament depenen d'algun factor com els hàbits.
Mesures: Nivell de LDL, tenint en compte un valor de eficàcia com pot ser la variació en un mateix pacient.
Dimensions: Hàbit, mes, setmana (L'evolució es podrà observar millor si agreguen a nivell de setmana.)

Hi ha diferencies en el resultat d'un tractament segons el lloc geogràfic del pacient?
Mesures: Nivell de LDL tenint en compte un valor de eficàcia com pot ser la variació en un mateix pacient.
Dimensions: tractament, ccaa i ciutat.

Hi ha algun període de l'any on el tractament sigui mes o menys efectiu.
Mesures: Nivell de LDL tenint en compte un valor de eficàcia com pot ser la variació en un mateix pacient.
Dimensions: mes (es considera agrupació de data suficient per veure un periode), tractament.

2.1.1 Disseny de les proves

Durant la fase de implementació dels processos ETL s'han realitzat una sèrie de proves per assegurar la qualitat de la execució de les carregues. A continuació es presenta la relació de proves realitzades durant la fase corresponent.

#	1
Descripció	Execució de la carregar quant el fitxer existeix
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb èxit i l'operador ha de rebre un mail dient que ha acabat OK
Resultat de la prova	OK

#	2
Descripció	Execució de la carregar quant el fitxer no existeix
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb ERROR i l'operador ha de rebre un mail dient que ha acabat KO
Resultat de la prova	OK

#	3
Descripció	Execució de la carregar quant la connexió a la BBDD ha caigut
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb ERROR i l'operador ha de rebre un mail dient que ha acabat KO
Resultat de la prova	OK

#	4
Descripció	Carregar de pacients duplicats
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb èxit, el pacient solament es guardarà a la taula dw_patients una vegada.
Resultat de la prova	OK

#	5
Descripció	Carregar de hàbits duplicats
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb èxit, el hàbit solament es guardarà a la taula dw_habits una vegada.
Resultat de la prova	OK

#	6

Descripció	Carregar de tractament duplicats
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb èxit, el tractament solament es guardarà a la taula dw_hreatments una vegada.
Resultat de la prova	OK

#	7
Descripció	Execució de la carregar quant el fitxer no existeix
Mòdul	ETL
Resultat esperat	El job ha de finalitzar amb èxit i l'operador ha de rebre un mail dient que ha acabat OK
Resultat de la prova	OK

2.2. Descripció del producte

Entorn d'execució de les carregues.

S'ha fet servir el sistema *PENTAHO DATA INTEGRATION (PDI)* per a la automatització dels processo ETL de la solució. El sistema permet gestionar la execució de tots els jobs del sistema i programarlos per mitjans del sistema escollit.

Per mitjans de la eina "kitchen" del sistema *PDI* i de crides com aquesta es pot executar els jobs del sistema. En aquest cas el nostre producte solament disposa d'un.

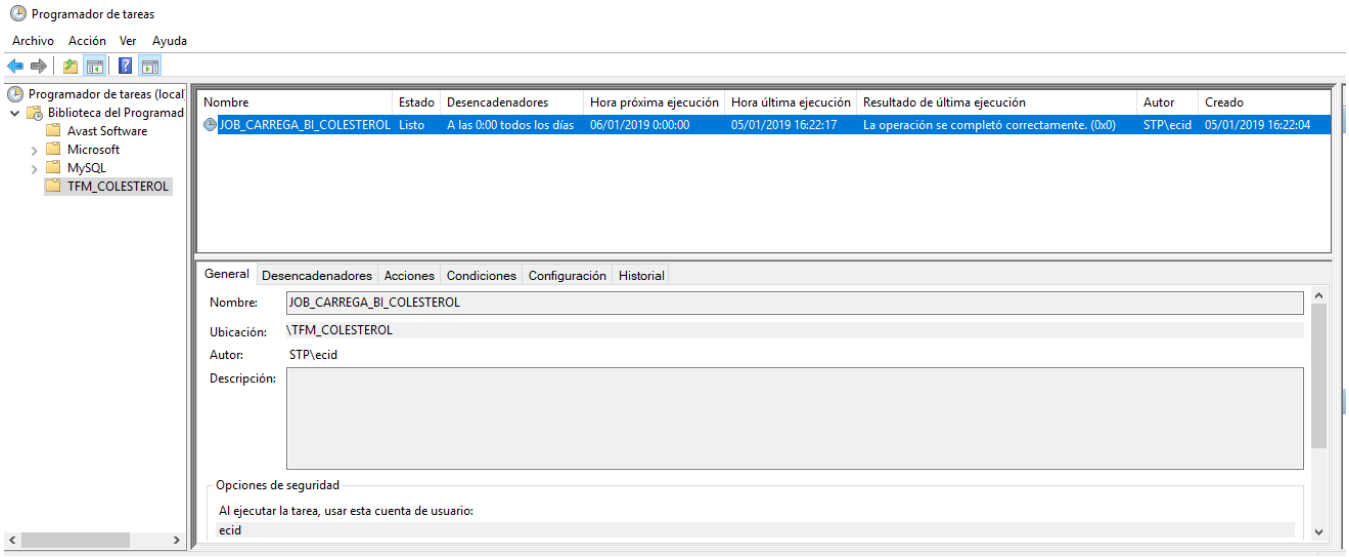
```
TO\JOB_CARREGA_BI_COLESTEROL.kjb"  "-param:CU
ENTA=PentahoBot" -logfile="c:\Users\ecid.STP\Dropbox\03 - UOC\MASTER\TFM\07-PRODU
CTO\LOG_JOB_CARREGA_BI_COLESTEROL.log"
```

Il·lustració 8 - Captura de la comanda del job amb kettle

La automatització d'aquests processos es pot fer per mitjans de diferents eines com per exemple:

- La eina *crontab* en sistemes Linux.
- La eina del *programador de tasques* pròpia de sistemes Windows.
- Eines de automatització especialitzades com "*Jenkins*", *IFTTT*,...

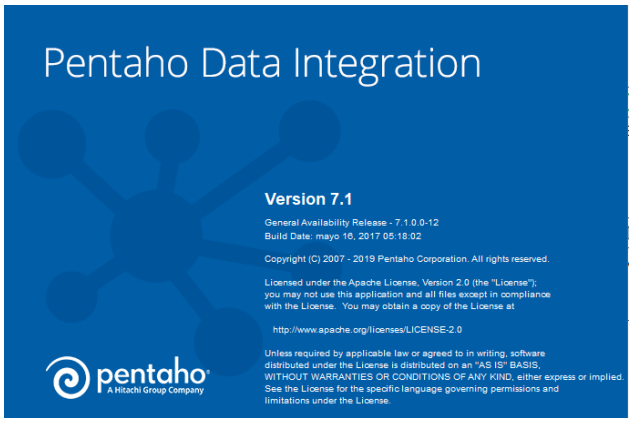
En el meu cas per simplificar he fet ser servir el programador de tasques de Windows. Primer s'ha creat un fitxer .bat per la execució de la comanda i després la tasca corresponent.



Il·lustració 9 - Captura de la tasca programada a windows

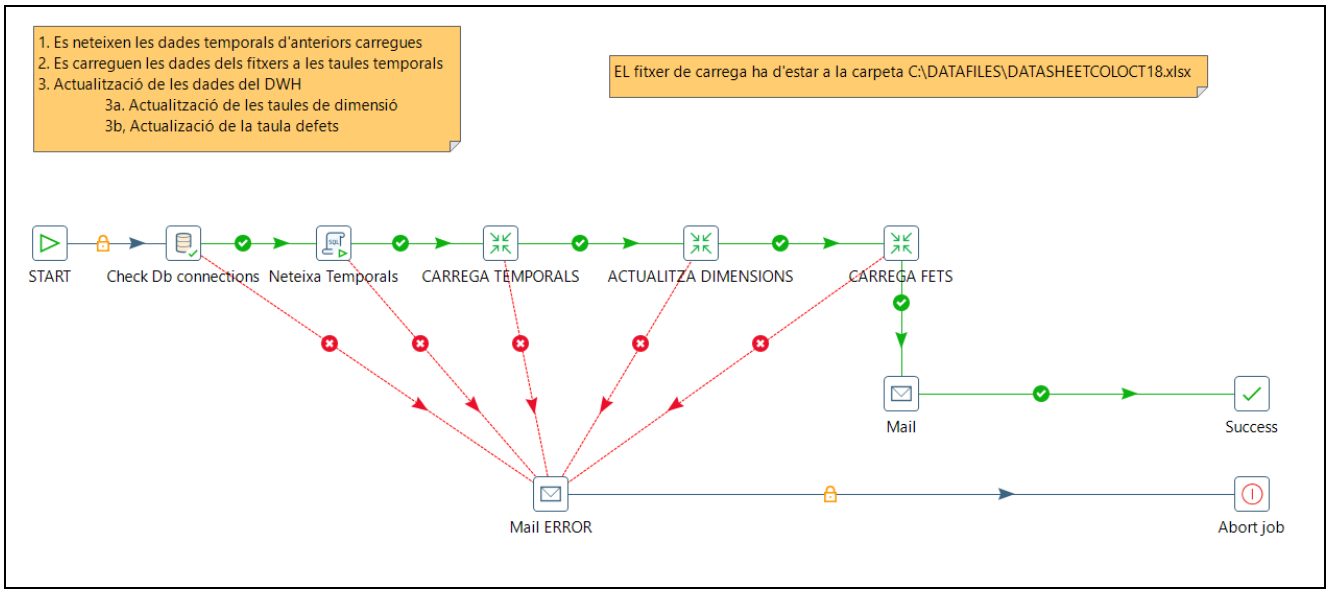
Processos de carrega de les dades.

Per mitjans de la eina Spoon del mateix entorn de PDI hem dissenyat els jobs i transformacions necessàries per al sistema ETL.



Il·lustració 10 - About de la eina spoon

El sistema disposa d'un job que gestiona tot el fluxe de la extracció de dades, transformació i carrega a la BBDD del datawarehouse.



Il·lustració 11 - Captura del job principal

El sistema està configurat per gestionar alarmes per mail als operadors corresponents tant si hi ha problemes en la execució con si finalitza correctament.

SE HA PRODUCIDO UN ERROR EN EL LA CARGA DE BI COLESTEROL

Job:

JobName : JOB_CARREGA_BI_COLESTEROL
Directory : /
JobEntry : Mail ERROR

Previous results:

Job entry Nr : 0
Errors : 0
Lines read : 0
Lines written : 0
Lines input : 0
Lines output : 0
Lines updated : 0
Lines rejected : 0
Script exist status : 0
Result : false

Path to this job entry:

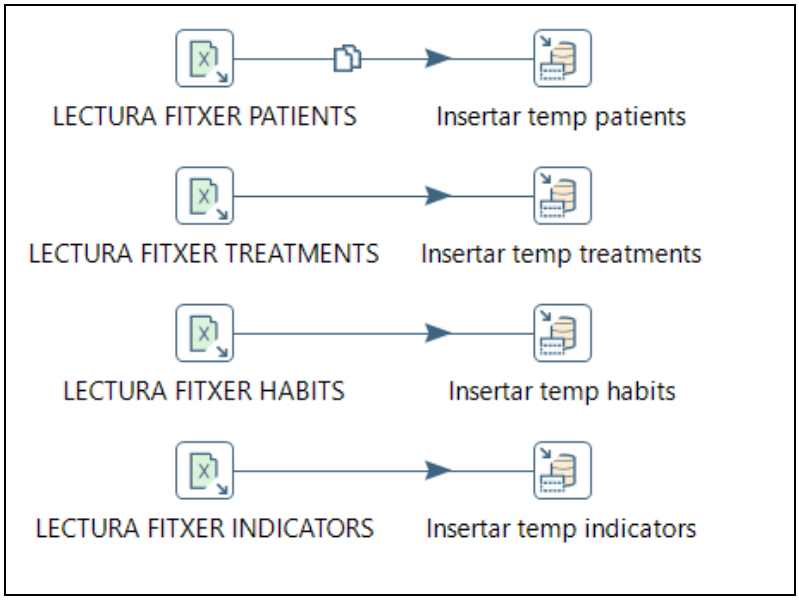
JOB_CARREGA_BI_COLESTEROL
JOB_CARREGA_BI_COLESTEROL : : start : Start of job execution (2018/11/18 14:23:59.467)
JOB_CARREGA_BI_COLESTEROL : : START : start : Start of job execution (2018/11/18 14:23:59.467)
JOB_CARREGA_BI_COLESTEROL : : START : [nr=0, errors=0, exit_status=0, result=true] : Job execution finished (2018/11/18 14:23:59.468)
JOB_CARREGA_BI_COLESTEROL : : Neteixa Temporals : Followed unconditional link : Start of job execution (2018/11/18 14:23:59.468)
JOB_CARREGA_BI_COLESTEROL : : Neteixa Temporals : [nr=0, errors=1, exit_status=0, result=false] : Job execution finished (2018/11/18 14:24:01.564)
JOB_CARREGA_BI_COLESTEROL : : Mail ERROR : Followed link after failure : Start of job execution (2018/11/18 14:24:01.564)

Il·lustració 12 - Alarma per mail d'una execució amb error del job

Para la implementació de las notificacions per Mail he utilitzat un servei de SMTP gratuit https://es.mailjet.com/

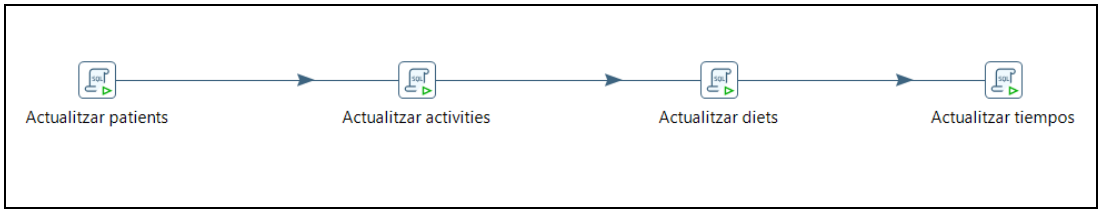
En primer lloc es fa el check de la connexió a la BBDD, després s'eliminen les dades de les taules temporals i comencen les transformacions del sistema.

La lectura del fitxer excel i la carrega de dades a les taules temporals:



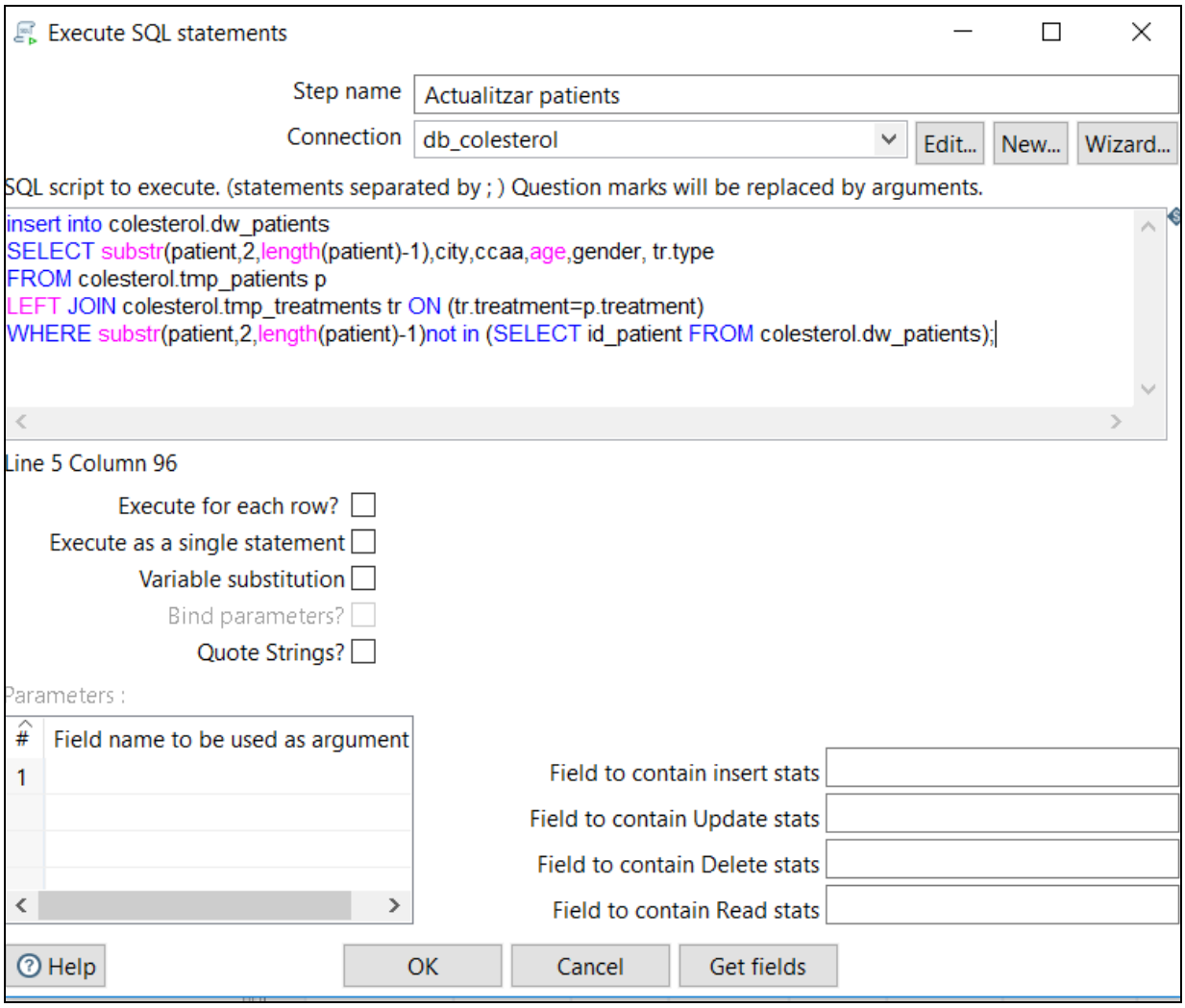
Il·lustració 13 - Transformació TRANS_CARREGA_TEMPORALS

Després s'executa la actualització de les dimensions:



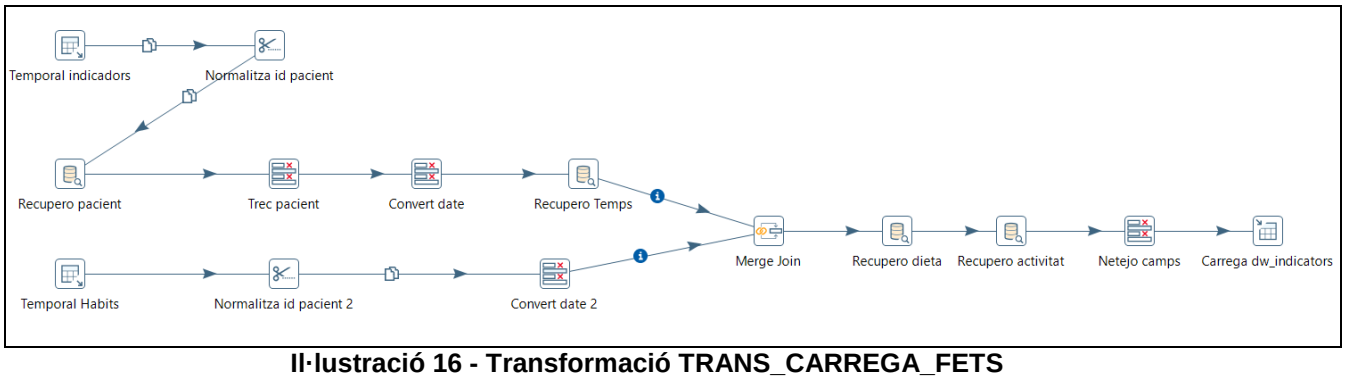
Il·lustració 14 - Transformació TRANS_ACTUALITZA_DIMENSIONS

En aquest cas la transformació disposa de diferents consultes SQL per actualitzar les taules de dimensions del DW.



Il·lustració 15 - Exemple de query per a la carrega de dimensions

Per ultim s'executa la transformació encarrgada de calcular els fets que s'analitzarán posteriorment.



Il·lustració 16 - Transformació TRANS_CARREGA_FETS

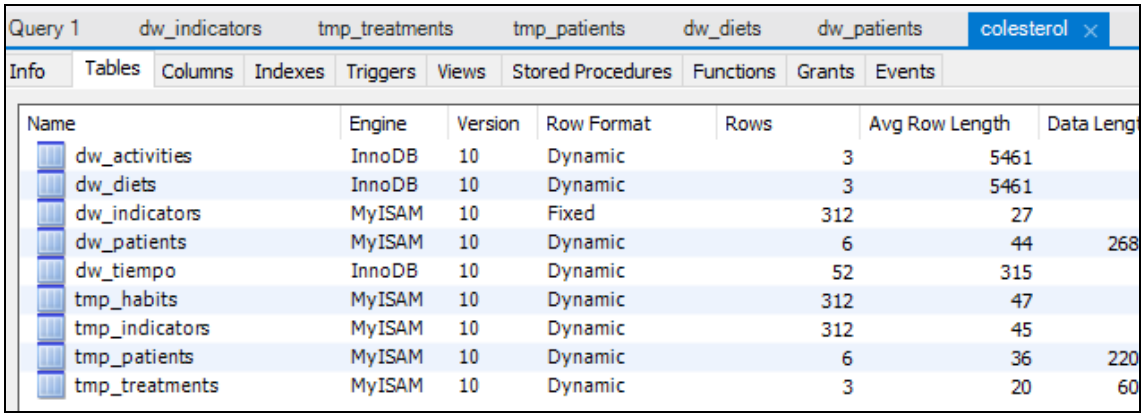
Una vegada finalitzada aquesta transformació i si el job no ha donat cap error finalitzarà amb èxit i notificarà a la persona corresponent la execució OK del job.

BBDD Datawarehouse.

El sistema gestor de base de dades escollit ha esta MYSQL per la experiència amb altres projectes similars i la facilitat que proporciona els diversos stacks que es poden trobar a la xarxa com wamp, xampp.

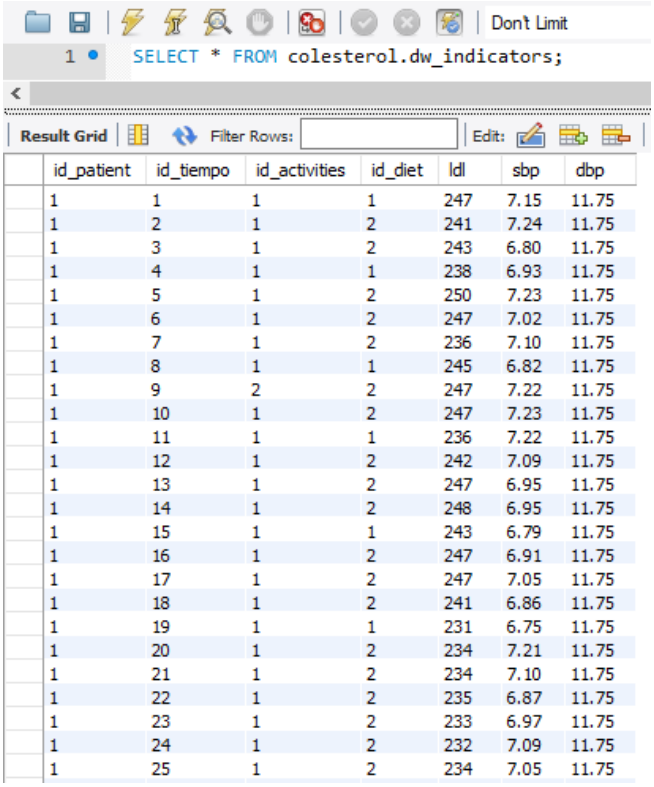
No obstant els sistema PENTAHO permet utilitzar qualsevol dels principals SGBD del mercat com Oracle, SQLServer, etc.

La estructura de taules resultant del disseny es la següent:



Name	Engine	Version	Row Format	Rows	Avg Row Length	Data Leng
dw_activities	InnoDB	10	Dynamic	3	5461	
dw_diets	InnoDB	10	Dynamic	3	5461	
dw_indicators	MyISAM	10	Fixed	312	27	
dw_patients	MyISAM	10	Dynamic	6	44	268
dw_tiempo	InnoDB	10	Dynamic	52	315	
tmp_habits	MyISAM	10	Dynamic	312	47	
tmp_indicators	MyISAM	10	Dynamic	312	45	
tmp_patients	MyISAM	10	Dynamic	6	36	220
tmp_treatments	MyISAM	10	Dynamic	3	20	60

Il·lustració 17 - Captura de les taules del DW

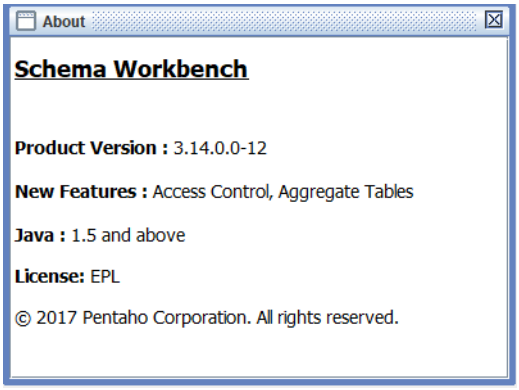


id_patient	id_tiempo	id_activities	id_diet	ldl	sbp	dbp
1	1	1	1	247	7.15	11.75
1	2	1	2	241	7.24	11.75
1	3	1	2	243	6.80	11.75
1	4	1	1	238	6.93	11.75
1	5	1	2	250	7.23	11.75
1	6	1	2	247	7.02	11.75
1	7	1	2	236	7.10	11.75
1	8	1	1	245	6.82	11.75
1	9	2	2	247	7.22	11.75
1	10	1	2	247	7.23	11.75
1	11	1	1	236	7.22	11.75
1	12	1	2	242	7.09	11.75
1	13	1	2	247	6.95	11.75
1	14	1	2	248	6.95	11.75
1	15	1	1	243	6.79	11.75
1	16	1	2	247	6.91	11.75
1	17	1	2	247	7.05	11.75
1	18	1	2	241	6.86	11.75
1	19	1	1	231	6.75	11.75
1	20	1	2	234	7.21	11.75
1	21	1	2	234	7.10	11.75
1	22	1	2	235	6.87	11.75
1	23	1	2	233	6.97	11.75
1	24	1	2	232	7.09	11.75
1	25	1	2	234	7.05	11.75

Il·lustració 18 - Exemple de consulta d'una de les taules

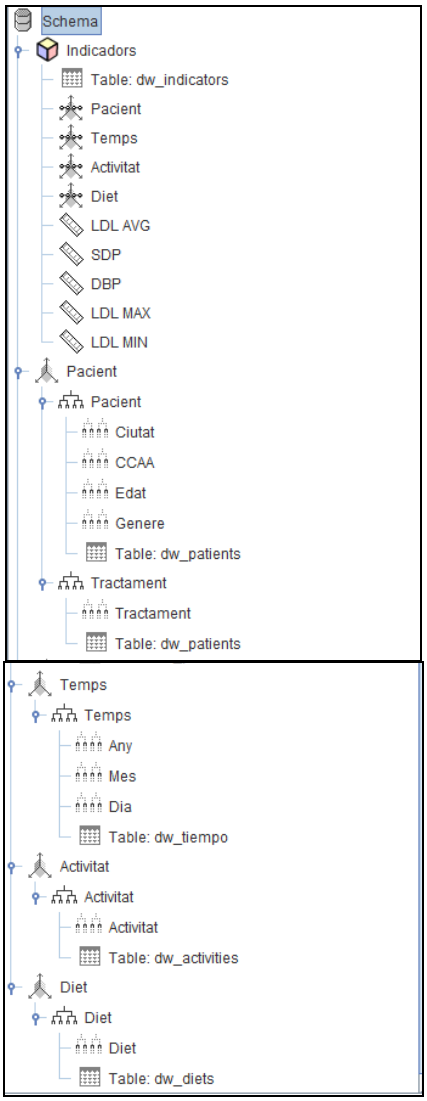
Estructura del cub.

Per poder generar els diferents llistats que ens ajudaran a fer l'estudi abans s'ha de crear una estructura del cub que farà servir l'entorn del *Pentaho Business Analytics*. Nosaltres hem fet servir la eina eina *schema workbench* per a generar aquesta estructura.



Il·lustració 19 - Captura de la finestra about de schema workbench

La estructura que he generat disposa de les dimensions Pacient, Temps, Activitat i Dieta. I com a fets hem utilitzats els indicadors de nivell de colesterol tenint en compte la mitjana, el màxim, el mínim, la mitjana de SDP y la de DBP.



Il·lustració 20 - Captura de la estructura del cub

El resultat del disseny del cub es un fitxer xml que es carregarà després a l'entorn corresponent per poder dissenyar els llistats.

Entorn de presentació.

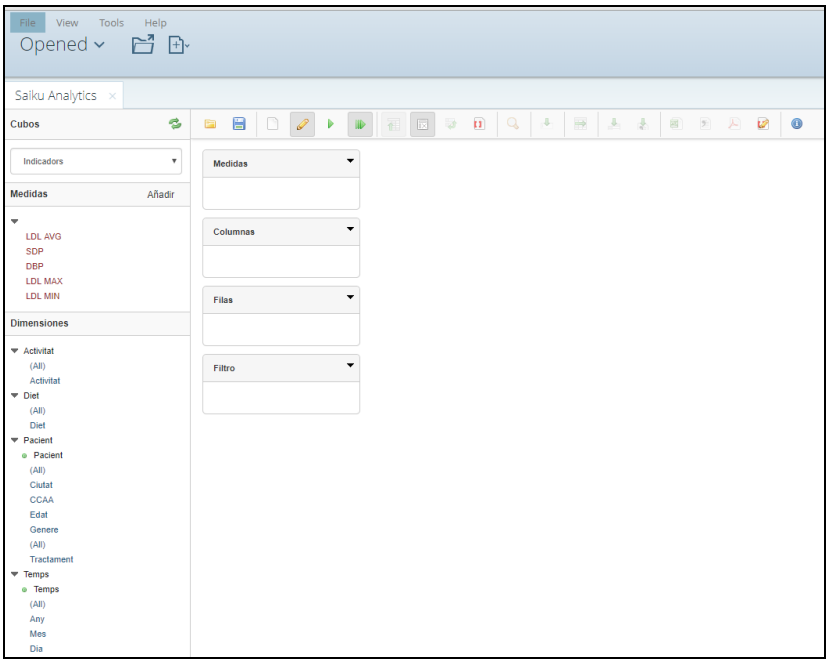
El sistema que gestiona el cub i la presentació de les dades en llistats es el servidor *Pentaho business analítics.* Es una aplicació web que s'executa sota un servidor tomcat.

Il·lustració 21 - Captura pantalla PENTAHO BA

L'entorn de Pentaho per si sol no pot gestionar els cubs dissenats amb la eina chema worskbech, sino que necesita de la instalació d'algún módul per interpretar-lo. En el meu cas he utilitzat el módul *SAIKU*, molt utilitzat en aquests tipus de arquitectura.

Il·lustració 22 - Captura pantalla d'inici de SAIKU

Aquesta eina serveix per poder fer a partir de la configuració d'un *datasource* que dona la connexió a la BBDD DW i la estructura del cub proporcionada, els corresponents llistats per a la realització de l'estudi de les dades de la investigació.



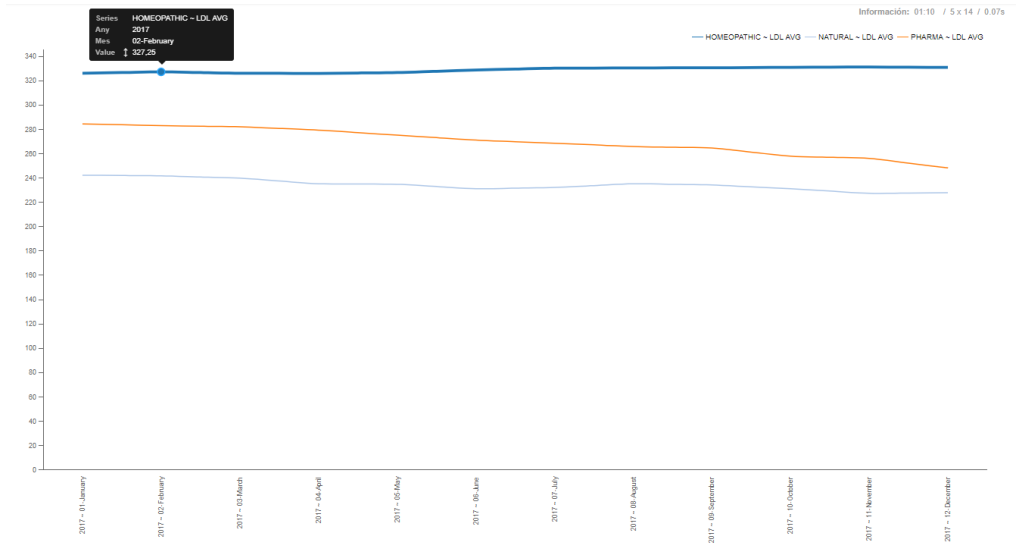
Il·lustració 23 - Captura dissenyador de reports - SAIKU

Llistats.

Els llistats que s'han dissenyat amb aquest entorn intenten resoldre les preguntes analítiques proposades als requeriments del projecte al apartat [1.3 Enfocament i mètode seguit.]

Llistat de la evolució per tractament.

Any	2017											
Mes	01-January	02-February	03-March	04-April	05-May	06-June	07-July	08-August	09-September	10-October	11-November	12-December
Tractament	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG
HOMEOPATHIC	326,1	327,25	326,125	326	326,7	328,75	330,3	330,5	330,625	331	331,25	330,875
NATURAL	242,3	241,75	239,875	235,25	234,8	231,125	232,2	235,25	234,25	231,1	227,375	227,875
PHARMA	284,5	283	282,125	279,375	275,2	271,125	268,6	265,875	264,625	257,9	256,125	248,25

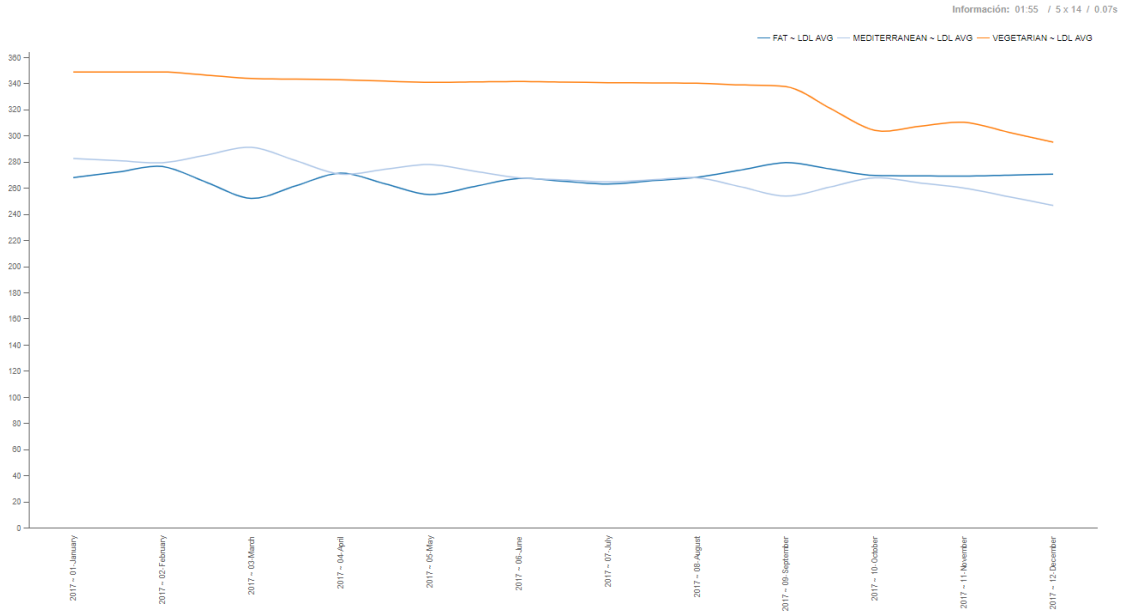


Il·lustració 24 - Gráfica de la evolució per tractament.

Segons les dades la evolució del tractament farmacologic durant 2017 sembla el millor dels 3, perque es el que mes variació té.

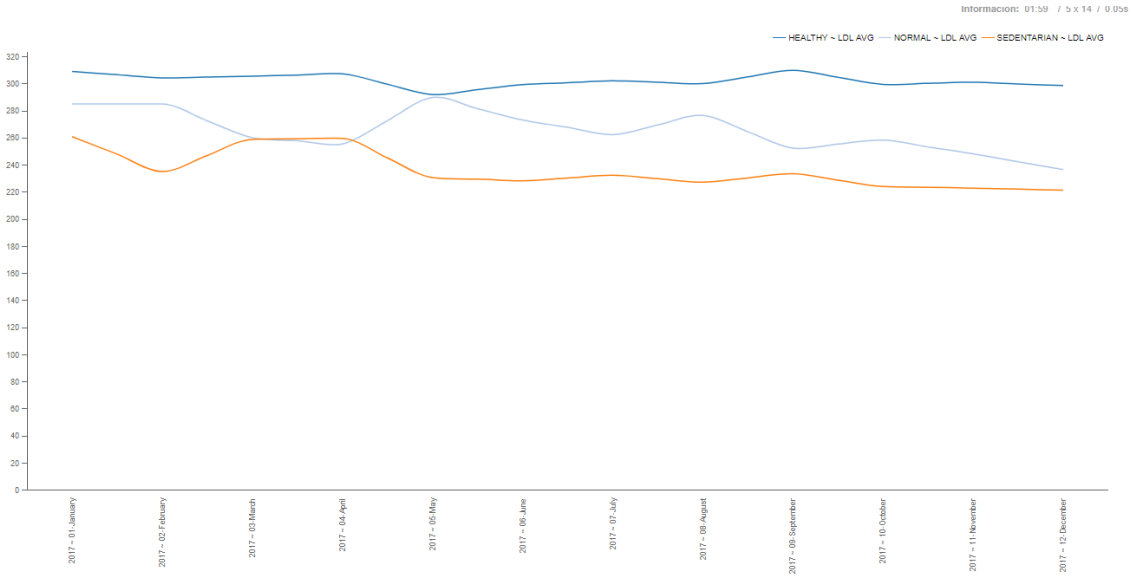
Llistat de la evolució pels hàbits (Dietes i activitats).

Any	2017											
Mes	01-January	02-February	03-March	04-April	05-May	06-June	07-July	08-August	09-September	10-October	11-November	12-December
Diet	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG
FAT	268,133	276,636	252,25	271,636	255,2	267,5	263,222	268,375	279,625	269,75	269,25	270,833
MEDITERRANEAN	282,818	279,546	291,357	270,8	278,188	268	264,875	268,077	254	268,056	260,167	246,9
VEGETARIAN	349	349	344	343	341	341,667	340,8	340,333	337,75	304,25	310,5	295,25



Il·lustració 25 - Gràfica evolució pels hàbits

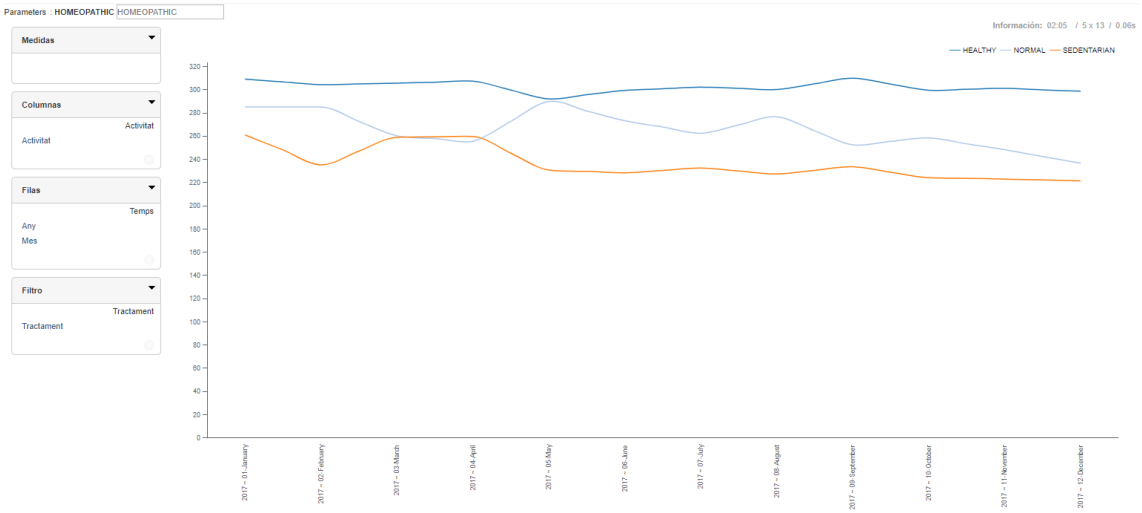
Any	2017											
Mes	01-January	02-February	03-March	04-April	05-May	06-June	07-July	08-August	09-September	10-October	11-November	12-December
Activitat	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG	LDL AVG
HEALTHY	309,143	304,333	305,667	307,364	292	299,417	302,214	300,182	310	299,6	301,167	298,692
NORMAL	285,133	285,091	260,286	255,625	290	273,286	262,5	276,75	252,5	258,5	248,333	236,667
SEDENTARIAN	261	235,25	259	259,8	230,833	228,4	232,5	227,4	233,667	224,2	223	221,5



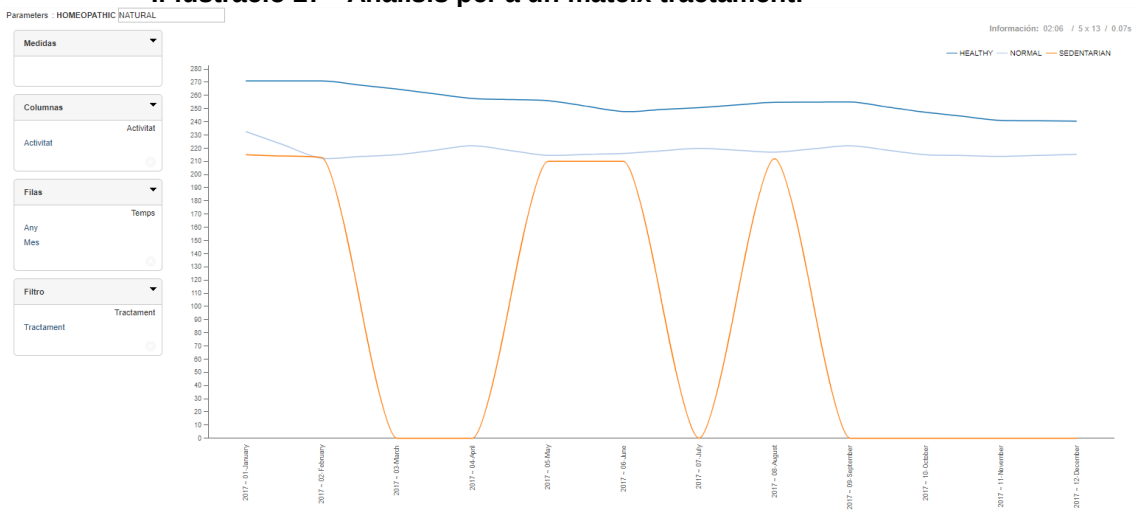
Il·lustració 26 - Gràfica evolució pels hàbits

Segons els gràfics sembla que els hàbits d'una dieta vegetariana i uns hàbits sans son els que millor resultat tenen a l'estudi.

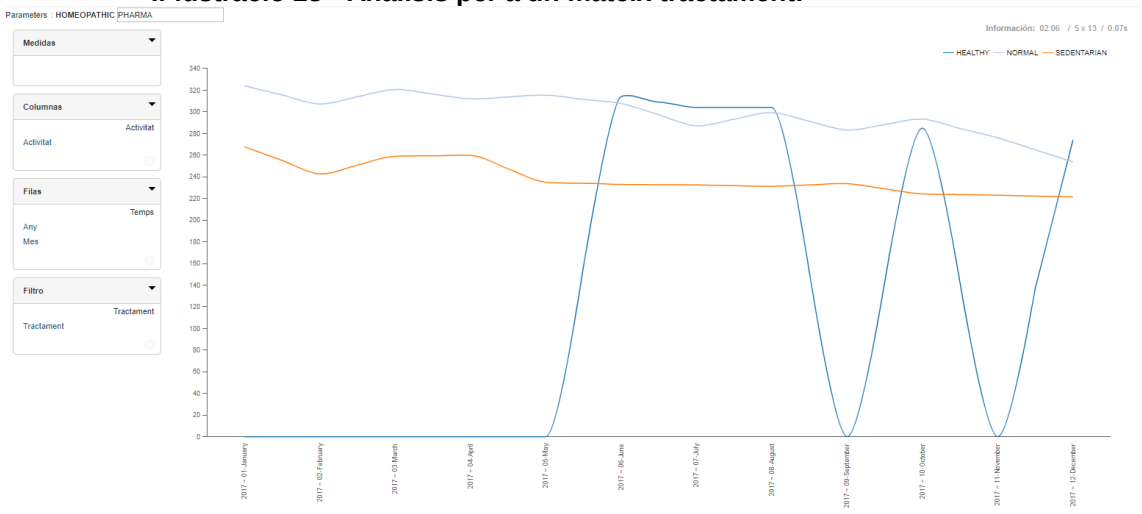
Anàlisis per a un mateix tractament.



Il·lustració 27 - Anàlisis per a un mateix tractament.



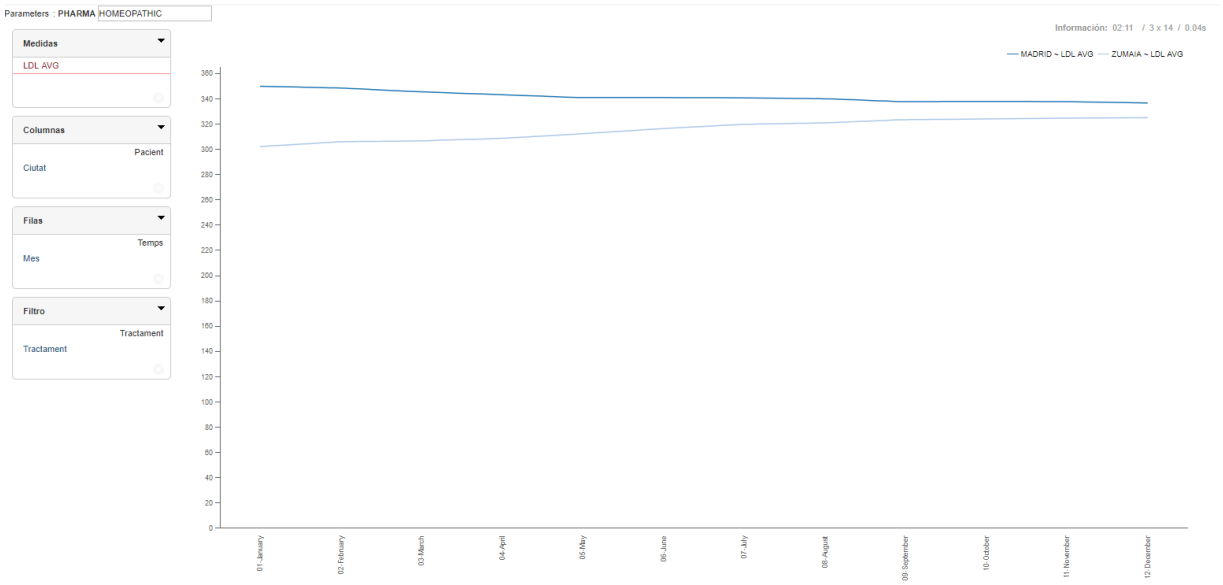
Il·lustració 28 - Anàlisis per a un mateix tractament.



Il·lustració 29 - Anàlisis per a un mateix tractament.

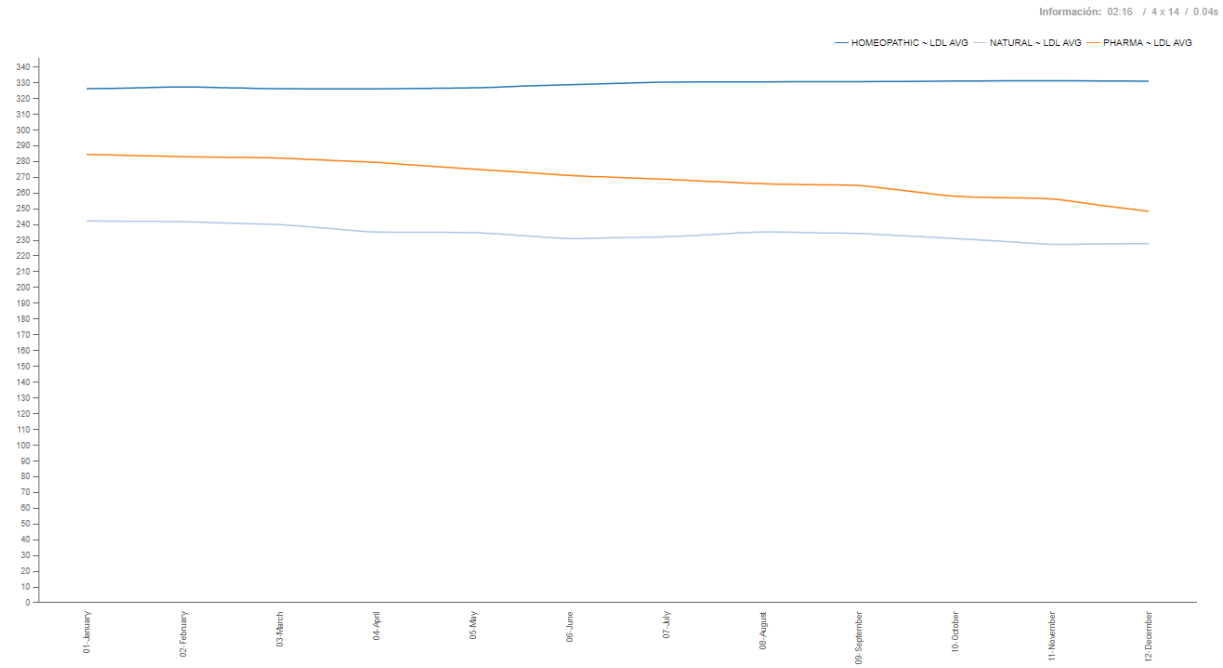
En el cas de la anàlisis per tractament hi ha certs hàbits que no disposen de dades complertes, s'hauria de fer una neteja de les dades o normalitzar-les per poder donar un anàlisis més exacte. El que si es pot veure es que la dieta sana amb un tractament natural es la que millor resultat obté.

Llistats del resultat segons el lloc geogràfic.



Il·lustració 30 - Gràfica segons el lloc geogràfic

Llistats dels resultat per període.



Il·lustració 31 - Gràfica segons el periode

A partir del resultat d'aquests llistats i gràfiques es podria fer l'estudi mes exacte de les dades de la investigació per poder respondre a les preguntes, però crec que el anàlisis no forma part del projecte sinó que ho faran posteriorment els investigadors amb la ajuda de la eina implementada.

3. Conclusions

Aquest treball m'ha servit per aprendre d'una manera global com es pot implementar un sistema BI amb l'ecosistema *PENTAHO* i amb la estructura pròpia de cubs.
La meva experiència en BI sempre havia estat amb eines que no necessitaven aquest tipus d'arquitectura com *POWER BI* o *Qlikview*.

En principi el resultat del producte crec que se serveix per complir amb els objectius principals, responent a totes les preguntes que tenen els investigadors quant expressen les seves necessitats. No obstant, crec que poden existir altres solucions de presentació que millorin la experiència d'usuari en quant a la facilitat per a crear nous llistats.

De cara a la meva planificació a nivell personal vaig ser massa positiu en la previsió de dedicació que podia donar a les tasques del treball. Com sempre la part de investigació i formació ha sigut la que més infravalorada estava i al final la que més ha ocupat durant la evolució del treball. No obstant he intentat seguir-la lo més acuradament. He anat actualitzant-la periòdicament per poder tenir una visió més realista del que quedava per finalitzar.

No solament he fet servir eines com el diagrama de *gantt* per gestionar la planificació, a un nivell més baix he fet servir *KANBAN* d'una manera més manual (Amb fulles a la paret i post-it's). Opino que aquest tipus de tècniques a vegades serveixen per tenir diferents visions de la planificació i guanyar concentració en el projecte en el que estàs endinsat.

Per falta de temps he deixat algunes característiques per properes versions com:

- Millorar la presentació de llistats
- Extreure mesures dels fets més complexes.

4. Glossari

- LDL – Low density lípids – lipoproteïnes de baixa densitat o colesterol dolent.
- HDL – High density lípids - lipoproteïnes d'alta densitat o colesterol bó.
- SBP - systolic blood pressure – Pressió arterial alta
- DBP - blood pressure diastòlic – Pressió arterial baixa
- BI – Business Inteligence – Intel·ligència de negoci
- PDI – PENTAHO DATA INTEGRATION
- PBA – PENTAHO BUSINESS ANALYTICS
- DW – Data warehouse – Sistema per enmagatçemar la informació. Normalment una base de dades.
- ETL – Extrac, transform and load – Sistema per extreure la informació, per exemple de fitxers, transformar-la i posteriorment carregar-la a una base de dades.

5. Bibliografia

[1] Web de la Endocrine society, Organització internacional mèdica en el camp de la endocrinologia i el metabollisme.

https://www.hormone.org/audiences/pacientes-y-cuidadores/preguntas-y-respuestas/2012/hyperlipidemia

[2] Article amb una petita definició dels sistemes BI

https://www.deustoformacion.com/blog/gestion-empresas/cuales-son-componentes-business-intelligence-big-data

[3] Article de com carregar csv a MYSQL desde UNIX.

https://ericlondon.com/2011/04/10/a-bash-shell-script-to-import-a-large-number-of-csv-files-into-mysql.html

[4] Web de Oracle data integrator

https://www.oracle.com/middleware/technologies/data-integrator.html

[5] Web oficial del producte de qlik

https://www.qlik.com/es-es/products/qlik-sense

[6] Web oficial de Microsoft flow

https://emea.flow.microsoft.com/es-es/

[7] Web oficial de Pentaho data Integration

https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html

[8] web oficial de Oracle DB

https://www.oracle.com/database/technologies/index.html

[9] Web oficial de Mysql

https://www.mysql.com/

[10] web oficial de Vertica

https://www.vertica.com/

[11] web oficial de Microsoft POWER BI

https://powerbi.microsoft.com/es-es/

[12] web oficial de Pentaho reporting

https://community.hitachivantara.com/docs/DOC-1009856-pentaho-reporting

[13] web oficial de Qlik

https://www.qlik.com/es-es/products/qlikview

Informació de esquema en estrella

https://es.wikipedia.org/wiki/Esquema_en_estrella - Visitada el 13/11/2018

Video Tutorial de la plataforma PENTAHO per part de la empresa STRATEBI

https://www.youtube.com/watch?v=IQEHd27CdX4

6. Annexos

- Fitxer amb la presentació utilitzada al vídeo. Ecidl_presentacio_–TFM.pdf
- Fitxer amb el diagrama de gantt de la planificació del projecte. PLAN_ecicl_v20181231.gan
- Carpeta comprimida amb el producte que inclou els fitxers del cub, dels processos ETL i exemples de llistats. annexos.ZIP