Magatzem de dades per l'anàlisi de la conciliació laboral i familiar o *work-life balance*

PR1 – Anàlisi i disseny del magatzem de dades

Marc Bracons Cucó

A partir de l'anàlisi del context del cas i de les fonts de dades disponible es demana dissenyar i proposar un magatzem de dades que ofereixi suport a l'anàlisi de la conciliació de la vida laboral i familiar o work-life balance.

Per a això, s'han de dur a terme els punts següents:

1) Anàlisi dels requisits

2) Anàlisi de totes les fonts de dades proporcionades

3) Anàlisi funcional

4) Disseny del model conceptual, lògic i físic del DW

1. Anàlisi dels requisits

Per aprofundir en l'anàlisi de la conciliació de la vida laboral i familiar mitjançant les fonts de dades proporcionades integrades en un DW. Aquestes són algunes de les preguntes que el sistema ha de poder respondre com a mínim.

- Percentatges de persones que treballen des de casa, ordenats per la data corresponent de menor a major.
- Mitjana d'hores treballades per país de la UE, comparada amb la Mitjana anual dels EUA.
- **Variació de l'equilibri entre la vida laboral i familiar en funció de l'edat dels fills:** Utilitzar la Mitjana i la mediana d'hores treballades per pares i mares, segmentades segons l'edat dels fills.
- **Comparació entre sexes en diferents sector productius:** Analitzar la Mitjana d'hores treballades i el percentatge de treball des de casa per sexe en cada sector econòmic.
- **Impacte del tipus de contracte en la conciliació laboral i familiar:** Comparar les hores treballades i el percentatge de treball a casa segons el tipus de contracte.

- **Anàlisi de la situació laboral i la paternitat/maternitat:** Correlacionar la situació laboral amb el nombre de fills, utilitzant coeficient de correlació.
- **Tendències temporals en la conciliació laboral-familiar per país:** Calculant la mitjana d'hores treballades i el percentatge de treball des de casa per país en un rang d'anys.
- **Diferències en la conciliació laboral i familiar segons l'edat dels treballadors:** Calcular la variància respecte a la mitjana en funció de l'edat.

2. Anàlisi de les fonts de dades

En aquest estudi tenim 10 tipus de fonts de dades diferents, a continuació se'n fa un resum de l'estructura així com comentaris sobre els punts d'interès de cadascun d'ells.

1) **lfsa_ewhun2_linear.csv.gz.** Mitjana d'hores treballades setmanals a la UE per país, sexe, edat, categoria professional i activitat econòmica.

Nom de camp	Descripció	Tipus	Exemple
DATAFLOW	Flux de dades	Text	ESTAT:LFSA_EWHUN2(1.0)
LAST UPDATE	Última actualització	Data i hora	15/02/23 23:00:00
freq	Freqüència	Text	A
nace_r2	Classificació d'activitats econòmiques a la Unió Europea	Text	A
wstatus	Estatus de treball o categoria laboral de la persona	Text	CFAM
worktime	Tipus d'horari laboral	Text	FT
age	Grup d'edat	Text	Y15-24
sex	Sexe	Text	F
unit	Unitat de mesura	Text	HR
geo	Codi geogràfic	Text	AT
TIME_PERIOD	Període de temps al qual corresponen les dades	Int	2017
OBS_VALUE	Valor observat	Int + Null	351
OBS_FLAG	Etiqueta per indicar condiciones de les dades observades	Text	bu

Total de registres: 480551

Comentaris:

- **DATAFLOW:** Té el mateix valor per a tots els registres.
- **LAST_UPDATE:** Tots els registres van ser actualitzats per últim cop el mateix dia.
- **freq:** Tots els registres tenen valor A (anual).
- **nace_r2:** Codi per agrupar les organitzacions en funció de les seves activitats comercials, per exemple, A pertany a "Agricultura, ramaderia, silvicultura i pesca". Es poden veure els seus possibles valors a la taula "ESTAT_NACE_R2_en.tsv**"**
- **wstatus:** Codi que indica el tipus d'activitat econòmica. EMP significa Employed persons, SAL Employees, etc. Es poden veure tots els possibles valors a la taula "ESTAT_WSTATUS_en.tsv".
- **worktime:** Codi que indica el tipus d'horari laboral. PT (Part Time), FT (full time), NRP (no response). Es poden veure tots els possibles valors a la taula "ESTAT_WORKTIME_en.tsv".
- **age:** Codi per establir grups d'edat. Y15-24, dels 15 als 24 anys. Es poden veure els possibles valors a "ESTAT_AGE_en.tsv".
- **sex:** Male o Female. Es poden veure tots els possibles valors a la taula "ESTAT_SEX_en.tsv"
- **unit:** HR (hour) en tots els registres
- **geo:** Codi geogràfic, en la majoria de casos son països però hi ha dos codi que agrupen 27 i 20 països respectivament (EU27_2020, EA20)
- **TIME_PERIOD:** Valor màxim 2021 i mínim 2017.
- **OBS_VALUE:** És un valor que indica el nombre de registres obtingut amb aquelles característiques, pot estar en blanc.
- **OBS_FLAG:** codi que indica informació referent a la mesura registrada, pot estar en blanc.

2) **lfst_hhwahchi_linear.csv.gz.** Percentatge de persones que treballen a casa per país, grup d'edat, nombre i edat de fills.

Nom de camp	Descripció	Tipus	Exemple
DATAFLOW	Flux de dades	Text	ESTAT:LFST_HHWAHCHI(1.0)
LAST UPDATE	Última actualització	Data i hora	15/02/23 23:00:00
freq	Freqüència	Text	A
sex	Sexe	Text	F
age	edat	Text	Y18-24
n_child	Nombre de fills	Text	2
agechild	Edat del fill	Text	Y_LT6
unit	Unitat de mesura	Text	PC
geo	Codi geogràfic	Text	SE
TIME_PERIOD	Període de temps al qual corresponen les dades	Int	2017
OBS_VALUE	Valor observat	Int + Null	3.6
OBS_FLAG	Etiqueta per indicar condiciones de les dades observades	Text	bu

Total de registres: 12883

Comentaris: És una taula molt semblant a l'anterior, només comentarem les diferències.

- **n_child:** registre on hi ha el nombre de fills. Es poden veure els diferents possibles valors a "ESTAT_N_CHILD_en.tsv"
- **agechild:** Codi que agrupa les edats dels fills. Y_LT6 menys de 6 anys.
- **unit:** en caquest cas el valor és PC, que vol dir percentatge.

3) **BLS_US_weeklyhours.xlsx.** Mitjana d'hores treballades setmanals als Estats Units. Dades obtingudes de US Department of Labor.

Nom de camp	Descripció	Tipus	Exemple
Year	Any	Int	2015
Period	Periode	Text	M01
Label	Etiqueta	Text	2015 Jan
Observation Value	Valor de l'observació	Float	34.5

Total de registres: 97

Comentaris: A la part superior del document hi ha una taula amb informació sobre el document, però no aporta registres.

4) **CountryList.json.** Conté els noms dels països en ordre alfabètic, agrupats per zona i amb el seu codi ISO 3166-1 alpha-2 en format JSON. L'estructura del fitxer és la següent:

Nom de camp	Descripció	Tipus	Exemple
name	Nom de país	Text	'Spain'
code	Codi	Text	'ES'
region	Regió	Text	'Southern Europe'

Total de registres: 79

5) **ESTAT_AGE_en.tsv.** Llista de trams d'edats segons classificació de l'Eurostat.

Nom de camp	Descripció	Tipus	Exemple
Etiqueta	Etiqueta	Text	LFD
Explicació	Explicació de l'etiqueta	Text	Late foetal death

Total de registres: 653

Comentaris: Aquesta taula aporta informació sobre la taula "fsa_ewhun2".

6) **ESTAT_N_CHILD_en.tsv.** Llista del nombre de fills segons classificació de l'Eurostat.

Nom de camp	Descripció	Tipus	Exemple
Etiqueta	Etiqueta	Text	GE1
Explicació	Explicació de l'etiqueta	Text	1 child or more

Total de registres: 19

Comentaris: Aquesta taula aporta informació sobre la taula "lfst_hhwahchi_linear.csv.gz".

7) **ESTAT_NACE_R2_en.tsv.** Llista d'activitats econòmiques segons classificació de l'Eurostat.

Nom de camp	Descripció	Tipus	Exemple
Etiqueta	Etiqueta	Text	A
Explicació	Explicació de l'etiqueta	Text	Agriculture, forestry and fishing

Total de registres: 1329

Comentaris: Comentaris: Aquesta taula aporta informació sobre la taula "fsa_ewhun2".

8) **ESTAT_SEX_en.tsv.** Llista de valors relatius al sexe de les persones que treballen segons classificació de l'Eurostat.

Nom de camp	Descripció	Tipus	Exemple
Etiqueta	Etiqueta	Text	M
Explicació	Explicació de l'etiqueta	Text	Males

Total de registres: 6

Comentaris: Comentaris: Aquesta taula aporta informació sobre les taula "fsa_ewhun2" i "lfst_hhwahchi_linear.csv.gz"

9) **ESTAT_WORKTIME_en.tsv.** Llista de tipus d'horaris laborals de les persones que treballen segons classificació de l'Eurostat.

Nom de camp	Descripció	Tipus	Exemple
Etiqueta	Etiqueta	Text	PC1-24
Explicació	Explicació de l'etiqueta	Text	From 1 to 24 percent of a full-time

Total de registres: 30

Comentaris: Aquesta taula aporta informació sobre la taula "fsa_ewhun2".

10) **ESTAT_WSTATUS_en.tsv.** Llista d'estats laborals segons classificació de l'Eurostat.

Nom de camp	Descripció	Tipus	Exemple
Etiqueta	Etiqueta	Text	EMP_W
Explicació	Explicació de l'etiqueta	Text	Persons employed or previously employed

Total de registres: 71

Comentaris: Aquesta taula aporta informació sobre la taula "fsa_ewhun2".

2.1 Estimació de volumetria

En els projectes de disseny de factoria d'informació corporativa hi ha una primera fase en la qual es fa una càrrega inicial i, a posteriori, una segona fase per fer les càrregues incrementals de les dades noves que van arribant. Una possible estimació del volum de dades del magatzem per a la càrrega inicial de les dades seria la següent:

Nom de fitxer	Registres	Valors	Dades
lfsa_ewhun2_linear.csv.gz	480551	13	6.247.163
fst_hhwahchi_linear.csv.gz	12883	12	154.596
BLS_US_weeklyhours.xlsx	97	4	388
CountryList.json	79	3	237
ESTAT_AGE_en.tsv	653	2	1306
ESTAT_N_CHILD_en.tsv	19	2	38
ESTAT_NACE_R2_en.tsv	1329	2	2658
ESTAT_SEX_en.tsv	6	2	12
ESTAT_WORKTIME_en.tsv	30	2	60
ESTAT_WSTATUS_en.tsv	71	2	142

En total hi ha **6.406.600** dades

3. Anàlisi funcional

A continuació es proposa el tipus d'arquitectura per a la factoria d'informació que s'adequa millor al projecte. Per a això, es consideren els requisits funcionals i s'estableix la prioritat entre exigible (E) i desitjable (D). En el context d'aquesta activitat, els requisits exigibles són aquells que es demanen en l'enunciat, mentre que els desitjables són els que complementen l'activitat.

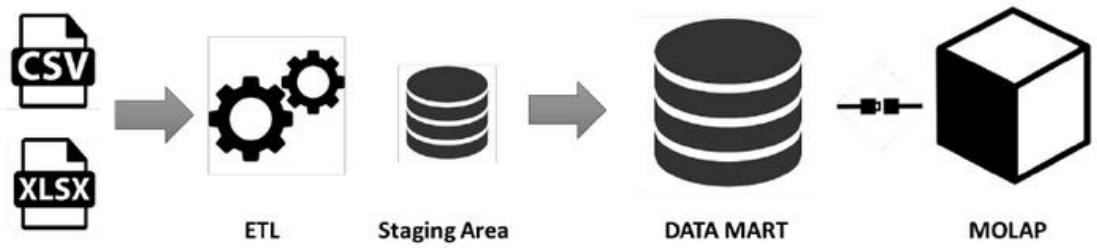
A més, en termes de l'escala de prioritats, s'assigna una prioritat de l'1 al 3, en què 1 és completament prioritari per a l'activitat i 3 és no prioritari.

A continuació es descriuen alguns dels requisits funcionals per al disseny d'una factoria d'informació per a l'organització tenint en compte les consideracions de l'enunciat:

#	Requeriment	Prioritat	Exigible / Desitjable
1	S'extreurà de forma adequada la informació de les fons de dades	1	E
2	Es crearà un DW	1	E
3	Es carregarà la informació sobre la conciliació laboral i familiar	1	E
4	Es crearà un model OLAP per a consultes dels usuaris	2	E
5	Es crearan els informes estàtics sol·licitats	2	E
6	Es redactarà un manual de càrrega de dades inicial i incremental	3	D

Per a triar l'arquitectura funcional s'han de tindre en compte els següents elements:

1. Les fonts de dades estan formades per 2 fitxers de text (csv), un full de càlcul (xlsx), 1 fitxer en format JSON i 6 fixters en format TSV.
2. L'arquitectura estarà formada per diversos elements
    - Staging area: Com en el nostre cas tenim múltiples fonts i de diferents tipus, és una bona idea carregar-les per consolidar la informació en una estructura intermèdia. Això és opcional però recomanable.
    - Data mart de la informació sobre la conciliació laboral i familiar o work-life balance. Tracta dades específiques.
    - MOLAP: a partir de la informació del data mart, es crearà un cub multidimensional per a les consultes dels usuaris.



Arquitectura FIC per a l'anàlisi sobre la conciliació laboral i familiar o work-life balance.

S'ha de tindre en compte que podrien existir altres requisits funcionals, tals com,

1. Creació de processos de qualitat de dades
2. Automatització de processos de càrregues incrementals
3. Seguretat i control d'accés
4. Monitoratge i registre d'activitats
5. Capacitat d'anàlisi predictiva i Machine Learning

4. Disseny del model conceptual, lògic i físic del magatzem de dades

4.1 Disseny conceptual

Per al correcte desenvolupament del DW, cal definir els fets (facts), les dimensions d'anàlisi (dimensions), les mètriques i els atributs que permetin tenir el nivell de granularitat suficient per a la presentació dels objectius.

- L'evolució de la conciliació de la vida laboral i familiar. Fa referència a la informació rellevant sobre indicadors relatius al work-life balance.

L'anàlisi de l'evolució d'indicadors relatius a conciliació de la vida laboral i familiar determina el disseny de les següents taules de fets:

Taula de fets	**Descripció**
FACT_PCT_EMPLOYEES_HOME	Anàlisi del percentatge de persones que treballen a casa

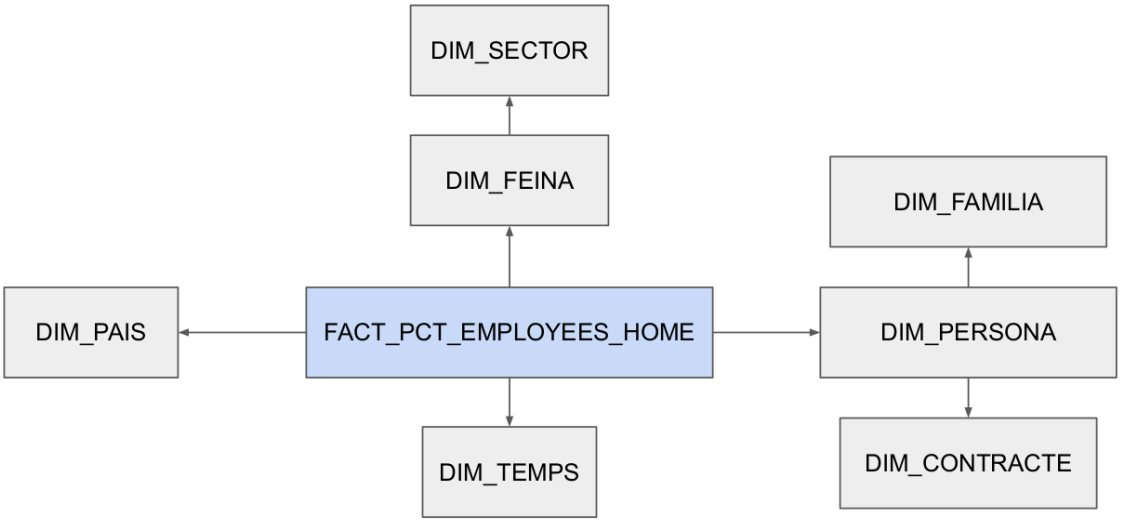
Les mètriques de la taula de fets FACT_PCT_EMPLOYEES_HOME és aquesta:

Mètriques	**Descripció**
OBS_VALUE_HOME	Percentatge de persones que treballen a casa

Aquestes mètriques poden ser analitzades des de les diferents dimensions:

Dimensions	**Descripció**
País	País de l'empresa on treballa cada persona
Temps	Temps en el que es fa l'anàlisi (mesos, anys)
Persona	Característiques de cada persona
Familia	Característiques familiars de cada persona
Feina	Característiques de la feina de cada persona
Sector econòmic	Sector econòmic al que pertany cada persona
Contracte	Tipus de contracte de cada persona

En la següent imatge es mostra el disseny del model per a la taula de fets FACT_PCT_EPLOYEES_HOME



Una altre possible taula de fets seria la que ens permeti analitzar com la quantitat d'hores treballades i la possibilitat de treballar des de casa canvien en funció del nombre i l'edat dels fills, cosa que podria indicar l'equilibri entre la vida laboral i familiar.

Taula de fets	**Descripció**
FACT_WORK_LIFE_BALANCE	Anàlisi relació feina-vida familiar

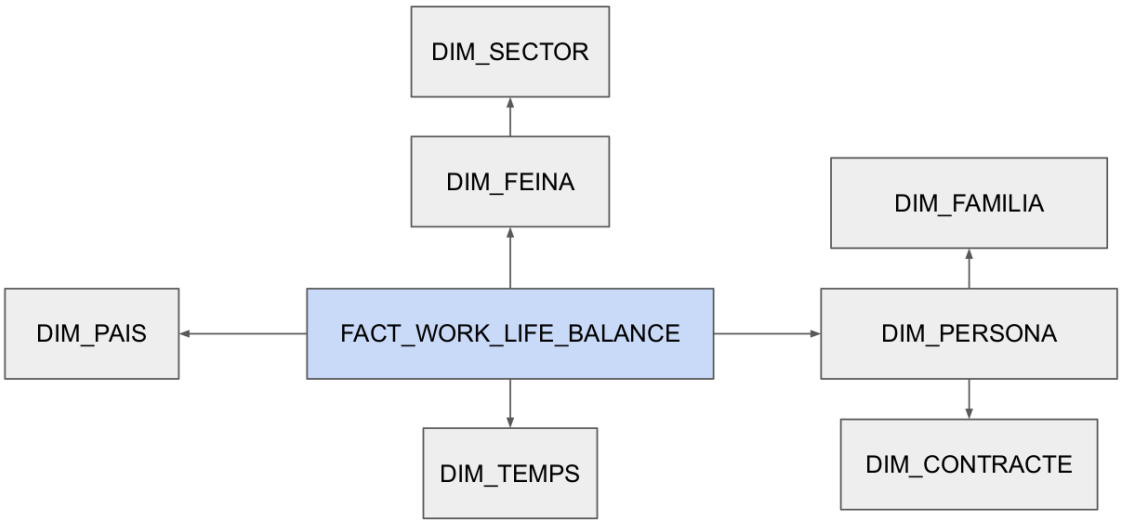
Les mètriques de la taula de fets FACT_WORK_LIFE_BALANCE és aquesta:

Mètriques	**Descripció**
OBS_VALUE_AVG	Mitjana d'hores treballades
OBS_VALUE_HOME	Percentatge de persones que treballen a casa

Aquestes mètriques poden ser analitzades des de les diferents dimensions:

Dimensions	**Descripció**
País	País de l'empresa on treballa cada persona
Temps	Temps en el que es fa l'anàlisi (mesos, anys)
Persona	Característiques de cada persona
Familia	Característiques familiars de cada persona
Feina	Característiques de la feina de cada persona
Sector econòmic	Sector econòmic al que pertany cada persona
Contracte	Tipus de contracte de cada persona

En la següent imatge es mostra el disseny del model per a la taula de fets FACT_WORK_LIFE_BALANCE



4.2 Disseny lògic

Una vegada obtingut el model conceptual del DW per a l'anàlisi d'indicadors de la conciliació de la vida laboral, passem a elaborar-ne el disseny lògic.

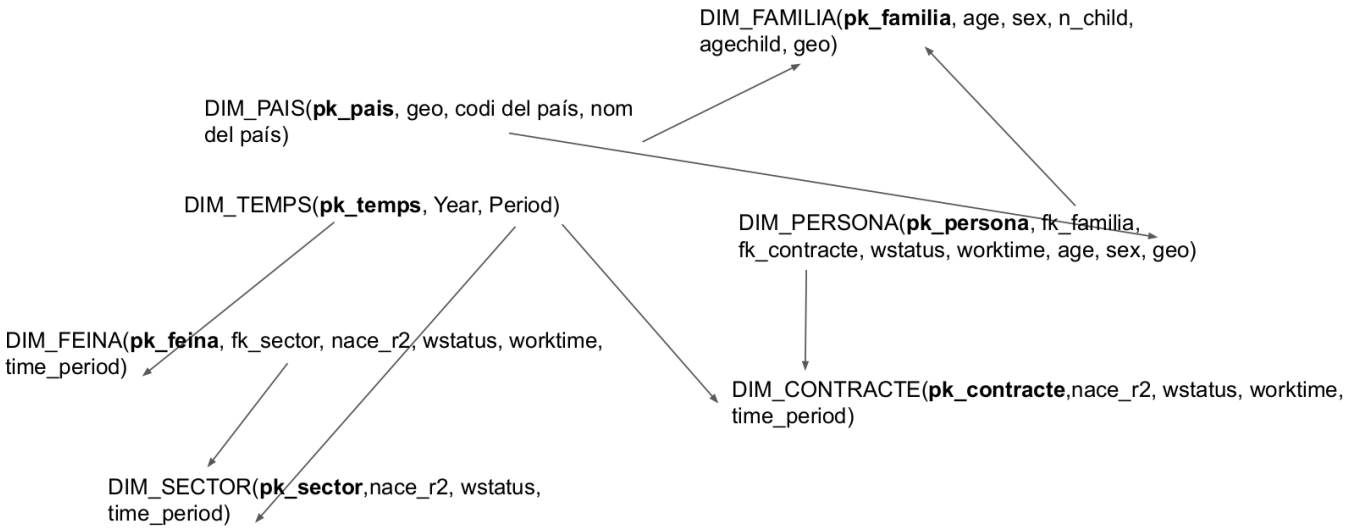
A continuació es mostra una taula amb alguna de les mètriques identificades en el disseny conceptual de la taula de fets FACT_PCT_EMPLOYEES_HOME:

Taula de fets	mètriques
FACT_PCT_EMPLOYEES_HOME	OBS_VALUE_HOME
FACT_WORK_LIFE_BALANCE	OBS_VALUE_AVG, OBS_VALUE_HOME

Després es detallen alguns dels atributs descriptors de les dimensions de cada fet:

Dimensions	**Atributs**
DIM_PAIS	Geo, codi del país, nom del país
DIM_TEMPS	Year, period
DIM_FEINA	nace_r2, wstatus, worktime, time_period
DIM_SECTOR	nace_r2, wstatus, time_period
DIM_FAMILIA	Age, sex, n_child, agechild, geo
DIM_PERSONA	Wstatus, worktime, age, sex, geo
DIM_CONTRACTE	nace_r2, wstatus, worktime, time_period

La representació visual del model lògic es pot veure en la imatge que hi ha a continuació



4.3 Disseny físic

Per al disseny físic del magatzem hem de tindre en compte diversos factors:

- El sistema gestor de bases de dades amb el qual treballarem implementarà d'una manera concreta els diferents elements del model lògic.

- Optimització del disseny físic al cas concret del nostre sistema gestor de bases de dades, obtenint així un correcte rendiment en les consultes.

- Revisió periòdic del disseny físic per assegurar que continua donant un correcte rendiment

Per a això, detallarem els tipus de dades de cada camp que formen part de les taules de fets i dimensions.

Dimensions

Les dimensions del model podran estar referenciades en les taules de fets utilitzant les seves claus primàries, o, en anglès, primary keys (PK). El model físic de les dimensions identificades és el següent:

DIM_PAIS: Conté les dades els daïsos

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_country(PK)	Numèric	8	1
Code	Text	8	ES
Country	Text	50	Spain
Region	Text	75	Souther Europe

DIM_TEMPS: Conté les dades temporals de les observacions

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_temps(PK)	Numèric	7	1267853
Year	Numèric	4	2015
Period	Text	3	M01

DIM_FEINA: Conté les dades professionals de la persona

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_feina(PK)	Numèric	12	101501000102
nace_r2	Text	21	A01
wstatus	Text	15	SELF
worktime	Text	7	H0
time_period	Numèric	4	2016

DIM_SECTOR: Conté les dades del sector en funció del sector comercial

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_sector(PK)	Numèric	10	1012141514
nace_r2	Text	21	A01
wstatus	Text	15	SELF
time_period	Numèric	4	2016

DIM_FAMILIA: Conté les dades familiars de la persona

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_familia(PK)	Numèric	9	101450123
age	Text	13	H0
sex	Text	4	DIFF
n_child	Text	3	GE1
agechild	Text	6	Y_GE12
geo	Text	9	EU27_2020

DIM_PERSONA: Conté les dades de la persona

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_persona(PK)	Numèric	10	1024782163
wstatus	Text	15	SELF
worktime	Text	7	H0
age	Text	13	H0
sex	Text	4	DIFF
geo	Text	9	EU27_2020

DIM_CONTRACTE: Conté les dades del contracte de la persona

Nom de camp	**Tipus**	**Mida**	**Exemple**
pk_contracte(pk)	Numèric	8	32456981
nace_r2	Text	21	A01
wstatus	Text	15	SELF
worktime	Text	7	H0

Per a donar valors a les claus primàries, s'ha volgut que siguin de tipus numèric, ja que això aporta eficiència en l'emmagatzematge, rapidesa d'ordenació, integritat referencial, auto incrementació i simplicitat i estandardització. A més, per temes de privacitat i evitar possibles cadenes d'errors, no és desitjable que a partir de la clau primària, una persona no vinculada a les dades, pugui deduir altres paràmetres com el nom del camp o els seus valors.

Pel que fa a la mida de la clau primària, s'han tingut en compte els camps vinculats a la dimensió i s'han sumat els dígits necessaris per a donar la informació de tots els camps. Per exemple, en la dimensió temps hi ha els camps Year i Period.

- Year és un enter de 4 digits (per exemple 2015)
- Period és un text de 3 caràcters (per exemple M01)
- Pk_temps tindrà mida 4+3=7

Això dona una mida superior a la necessària, però és una bona aproximació ja que no pugem ordres de magnitud.

Taula de fets

La composició del model físic de les taules de fets consistirà en la creació de taules els camps de les quals seran les mètriques, els atributs i els atributs referencials definits en el model conceptual i en el model lògic. Per crear els atributs referencials en les taules de fets, es defineixen com a claus foranes les primàries de les dimensions amb les quals estan relacionades, seguint el diagrama en estrella definit.

El model físic de les taules de fets del magatzem de dades per a l'anàlisi de la conciliació de la vida laboral i familiar està compost, per les taules següents:

FACT_PCT_EMPLOYEES_HOME. És la taula física que contindrà la informació que permetrà fer l'anàlisi de la conciliació de la vida laboral i familiar. Tindrà els camps següents:

Nom camp	**Tipus**	**Mida**	**Exemple**
pk_EH_id(PK)	Numèric	4	30
fk_pais	Numèric	8	1
fk_temps	Numèric	6	482
fk_feina	Numèric	12	167326485
fk_sector	Numèric	10	4783
fk_familia	Numèric	9	105258
fk_persona	Numèric	10	962574
fk_contracte	Numèric	8	253365
obs_value_home	Numèric	3	4

FFACT_WORK_LIFE_BALANCE. És la taula física que contindrà la informació que permetrà fer l'anàlisi de la relació entre la feina i la família. Tindrà els camps següents:

Nom camp	**Tipus**	**Mida**	**Exemple**
pk_EH_id(PK)	Numèric	4	30
fk_pais	Numèric	8	1
fk_temps	Numèric	6	25
fk_feina	Numèric	12	168485
fk_sector	Numèric	10	254783
fk_familia	Numèric	9	1058
fk_persona	Numèric	10	942574
fk_contracte	Numèric	8	25365
obs_value_home	Numèric	3	4
obs_value_avg	Numèric	4	36.9