

LAURA REDAPI

Consortium
GARR

THE ITALIAN
EDUCATION
& RESEARCH
NETWORK



UNIVERSITÀ
DEGLI STUDI
FIRENZE



HEPMED: data mining in High Energy Physics and MEDicine

GIORNATA DI INCONTRO
BORSE DI STUDIO GARR
"ORIO CARLINI"
6 DICEMBRE 2018 ROMA

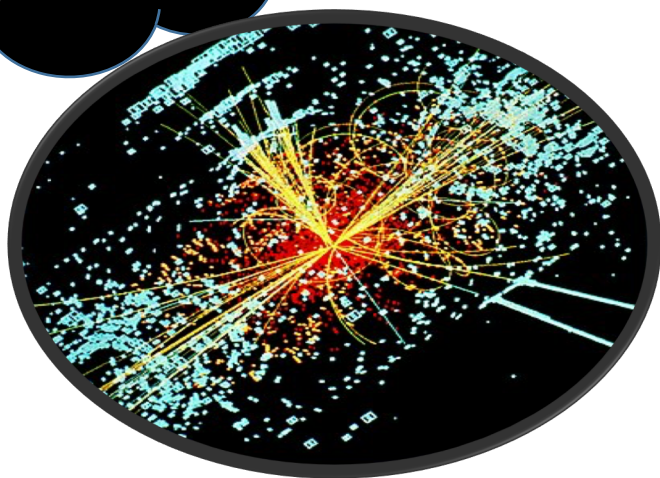
Roma

BORSISTI DAY 2018



Sviluppo di un modo efficiente di navigare grandi quantità di dati sfruttando tecniche di data warehousing, con applicazione diretta su database di **fisica delle alte energie** e di ambito **medico**

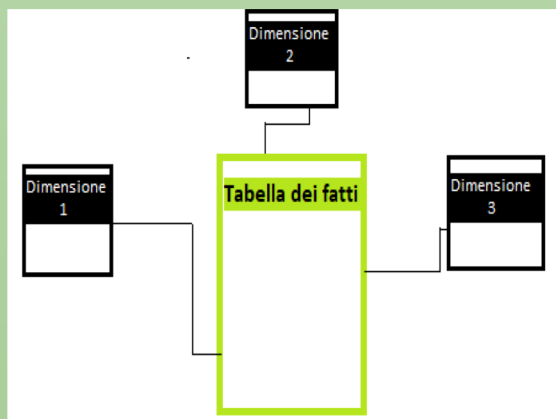
Abbiamo a disposizione dati di misure fatte dagli anni '70 ad oggi potrebbe esserci sfuggito qualcosa?



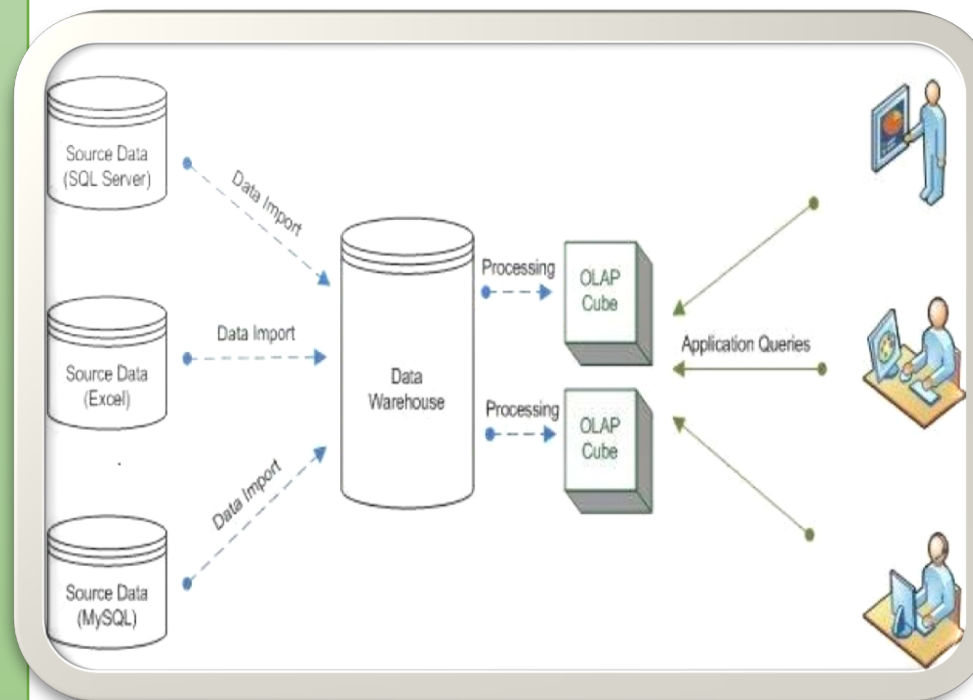
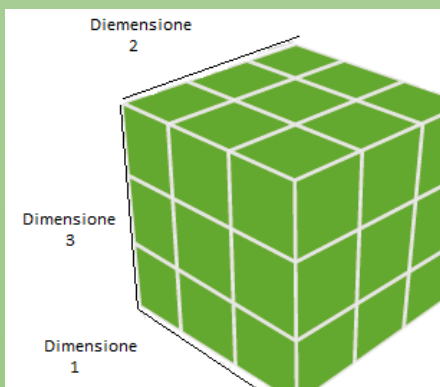
Correlando dati di esami medici diversi, potrei forse fare diagnosi più sicure?



- ✓ Un database multidimensionale e strutturato
Data Warehouse
-Star Schema



- ✓ Un modo efficiente di navigare e visualizzare i dati
OLAP tools (On-line Analytical Processing)
-Cubo multidimensionale





Strumenti a disposizione



HEPMED

si inserisce nel progetto dell'Università di Firenze MineHep:
-start point : **Hepdata** database di fisica delle alte energie.



- Repository open-access di dati di esperimenti di fisica delle alte energie dal 1970 (MySQL)



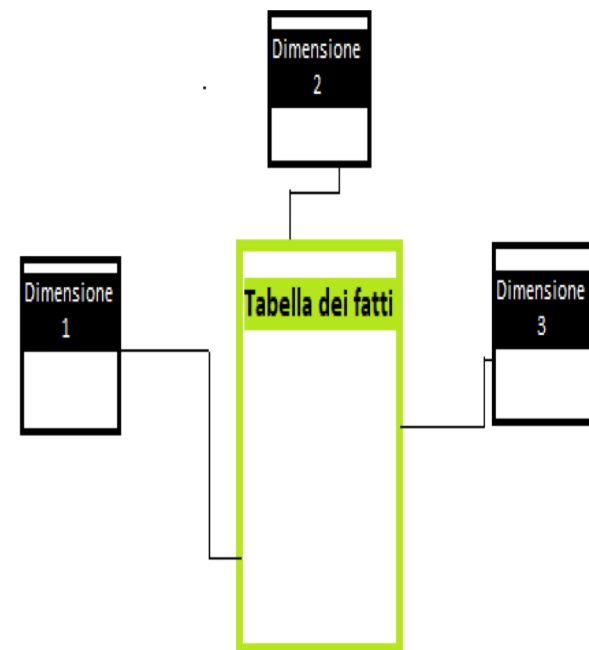
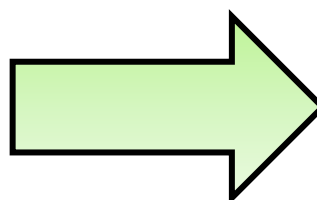
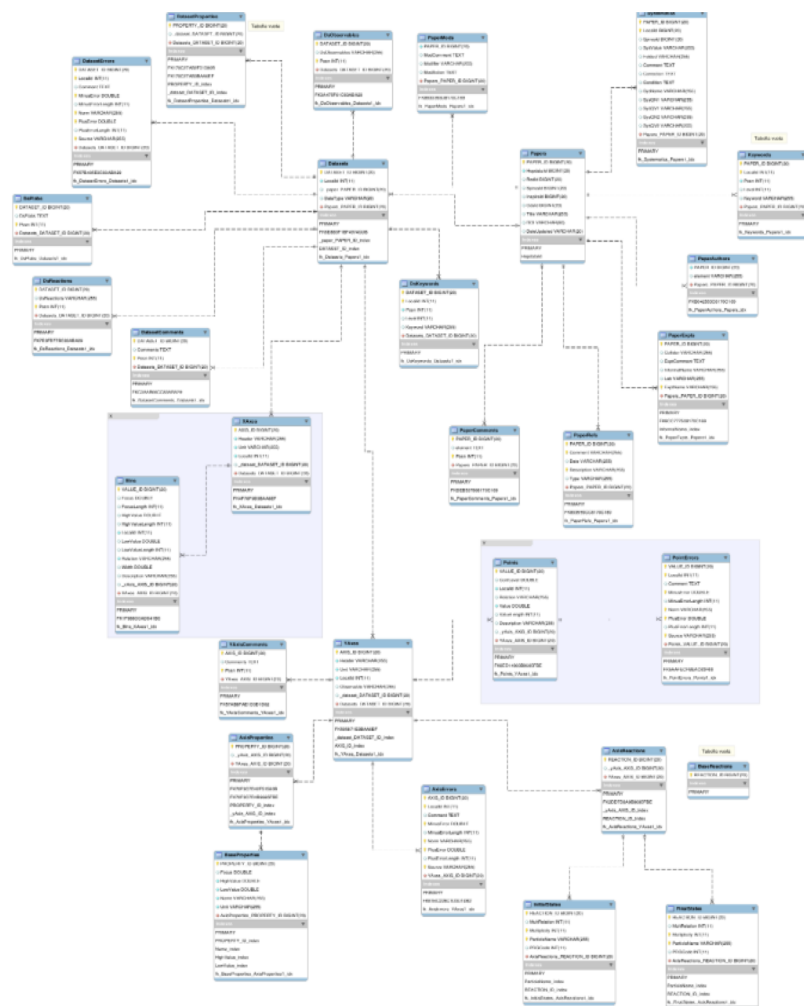
- supporta la lettura e il caricamento di database MySQL per la costruzione di Star Schema e di una dashboard per strutturare query sfruttando tecniche OLAP



GIORNATA DI INCONTRO BORSE DI STUDIO GARR “ORIO CARLINI” GIOVEDI' 6 DICEMBRE 2018 - ROMA

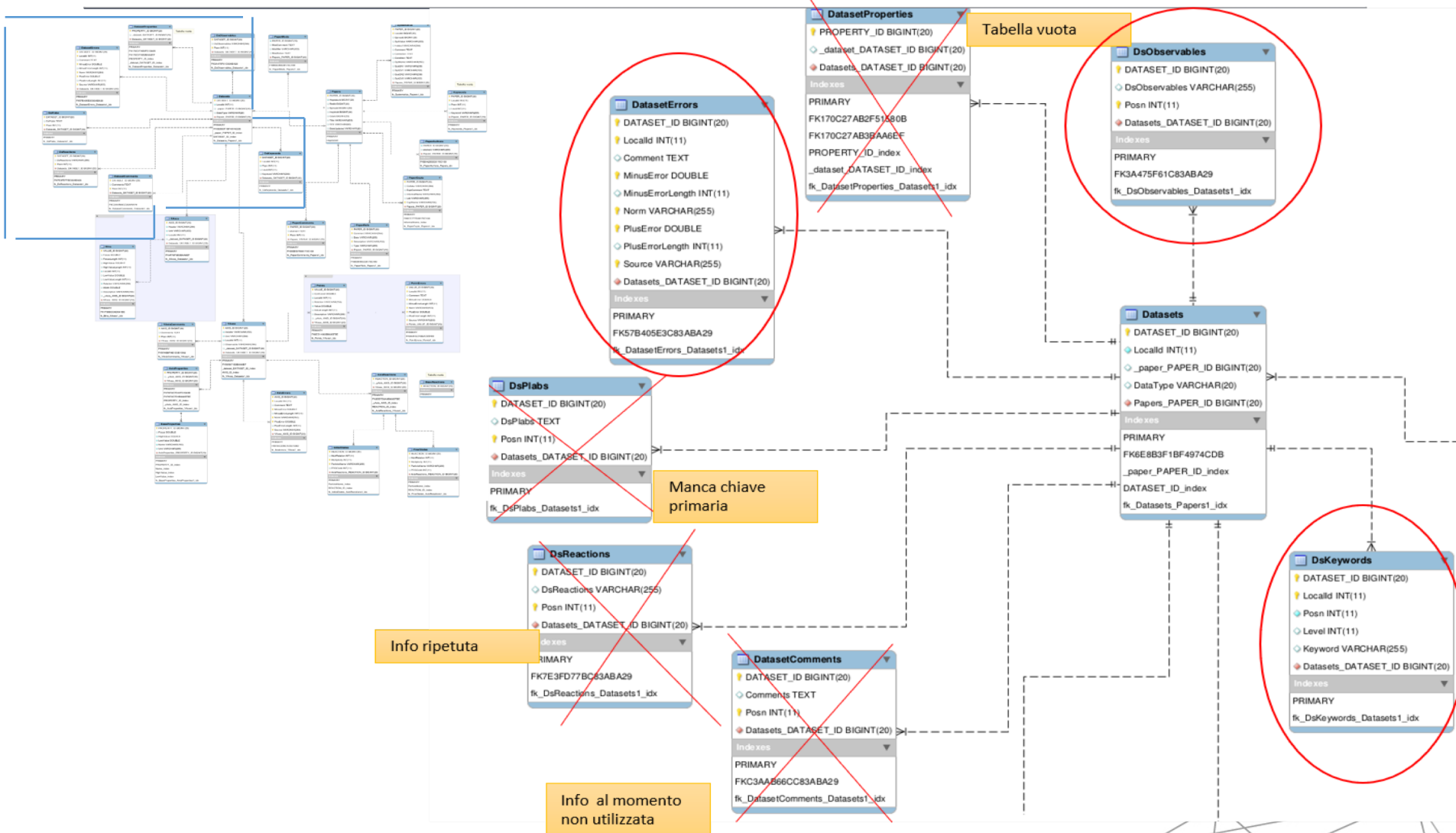


Hepdata





Hepdata

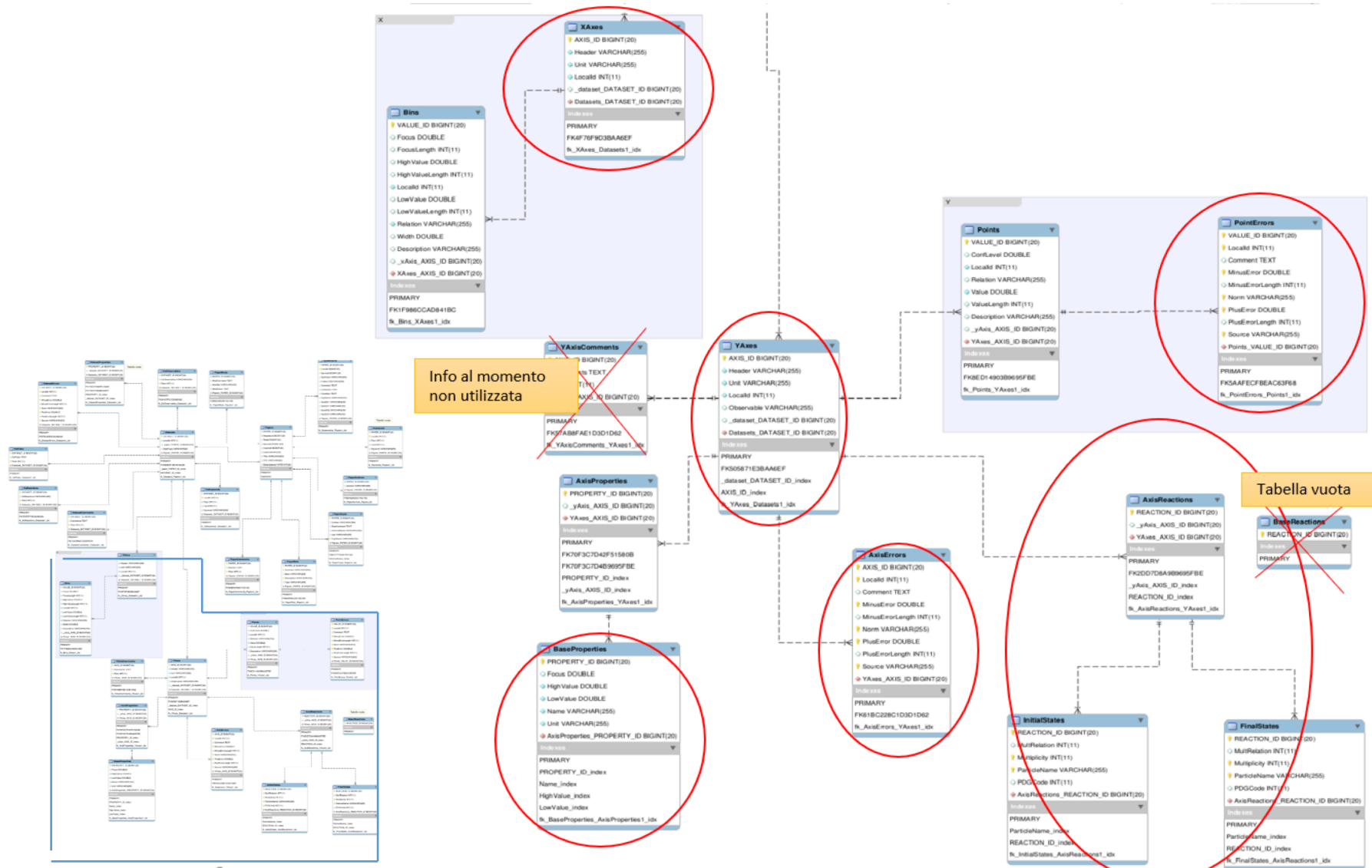






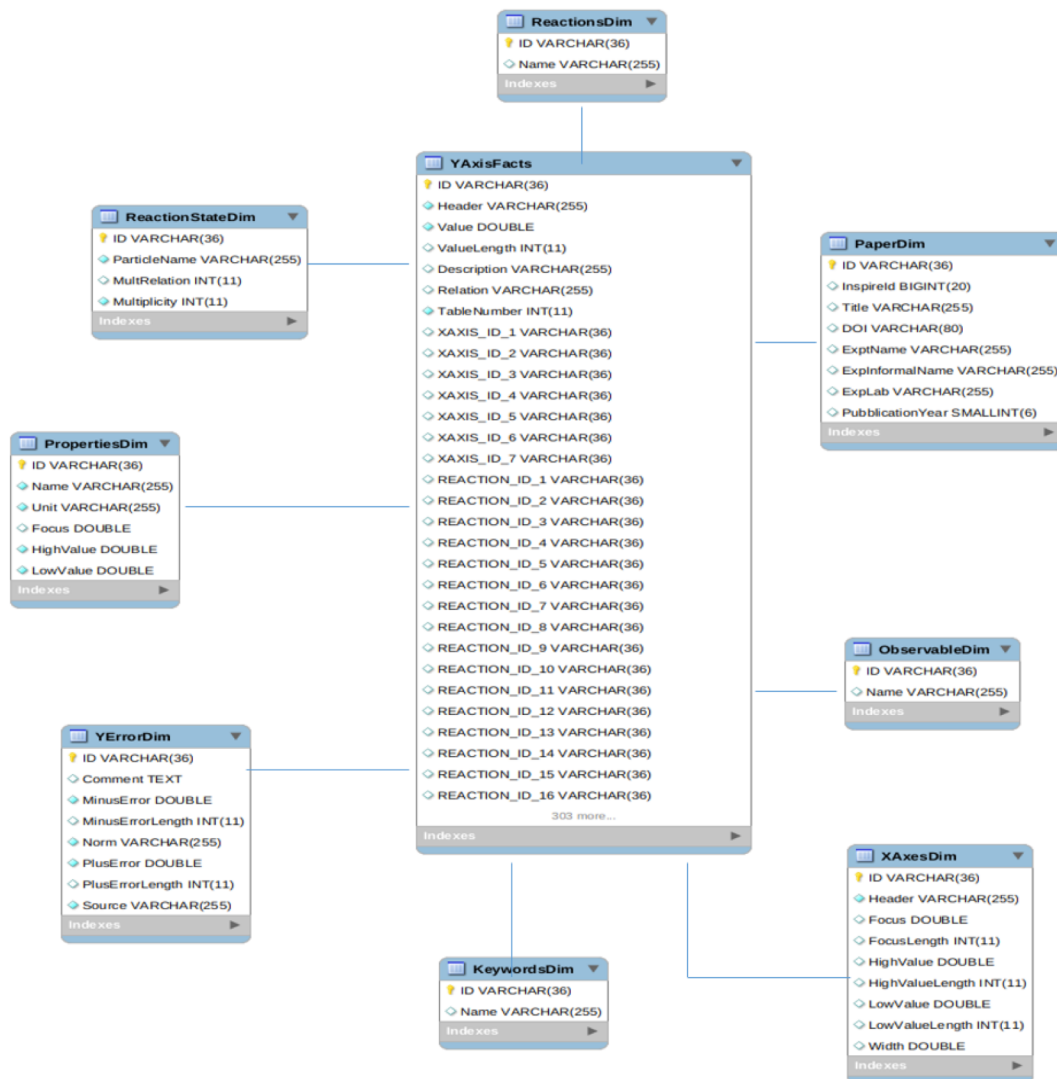
Hepdata

GIORNATA DI INCONTRO BORSE DI STUDIO GARR "ORIO CARLINI" GIOVEDI' 6 DICEMBRE 2018 - ROMA





MineHep (diagramma a stella)

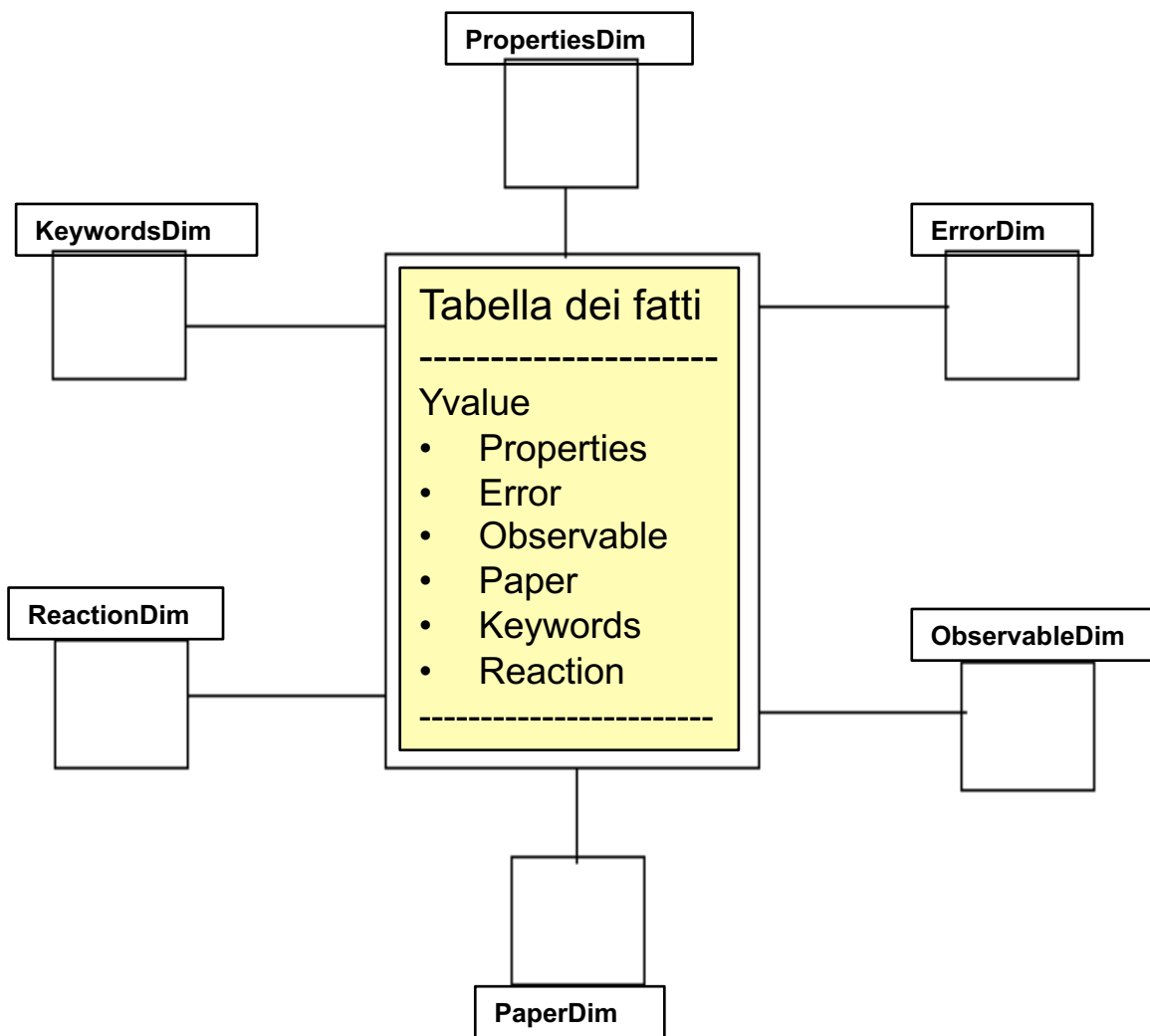


La **tabella dei fatti**, al centro, è l'unica a possedere collegamenti multipli, che si concretizzano attraverso identificatori univoci (campi chiavi) in essa contenuti, con le altre tabelle.

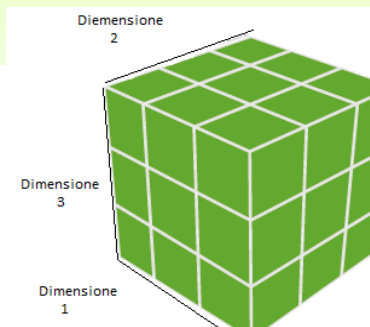
Le **dimensioni** hanno tutte un solo collegamento alla tabella centrale, con lo scopo di minimizzare il numero di join richiesti per ciascuna query



MineHep (diagramma a stella)



Una volta studiato il database di partenza lo scheletro del "nuovo" database da popolare avrà sempre questa struttura funzionale a qualsiasi interrogazione successiva anche complessa che verrà sviluppata tramite la tecnica OLAP





“Select” di controllo sul nuovo database **MineHep**

-- TEST: prendo tutte le misurazioni della tabella 1 che sono comprese tra [0.8, 2.0] $\sqrt{s}$ nella colonna HERAPDF

```
select distinct Value
from YAxisFacts yaf
inner join PaperDim ppd on ppd.ID = yaf.PAPER_ID
inner join XAxesDim xad on xad.ID = yaf.XAXIS_ID_1
where InspireId = 1118047 and TableNumber = 1
      and xad.Header = '$\sqrt{s}$'
      and yaf.Header = "HERAPDF"
      and xad.LowValue >= 0.8
      and xad.HighValue <= 2.0;
# Value
# 196
# 181
# 153
# 140
# 132
```

-- RESULT: OK

-- TEST: prendo tutte le misurazioni della tabella 1 della colonna $\sqrt{s}$ che hanno un errore SYS +/- 6

```
select yaf.Value
from YAxisFacts yaf
inner join PaperDim ppd on ppd.ID = yaf.PAPER_ID
inner join YErrorDim yed on yed.ID = yaf.POINT_ERROR_ID_1
where InspireId = 1118047
      and TableNumber = 1
      and Source = 'SYS'
      and Header= '$\sqrt{s}$'
      and (MinusError=-6 and PlusError=6);
# Value
# 156
# 136
```

-- RESULT: OK



PT(E)	> 35 GEV					
RE	P P → W ⁺ < E ⁺ NUE > X					
SQRT(S)	7000.0 GeV					
$ \eta $	$\mathcal{A}$	CT10	HERAPDF	MSTW	NNPDF	
0.0 - 0.2	102.0 ± 3.0 stat ± 5.0 sys	109.0 ± 5.0	106.0 $^{+4.0}_{-8.0}$	87.0 $^{+3.0}_{-5.0}$	107.0 ± 5.0	
0.2 - 0.4	111.0 ± 3.0 stat ± 5.0 sys	114.0 ± 5.0	110.0 $^{+4.0}_{-8.0}$	89.0 $^{+3.0}_{-5.0}$	110.0 ± 5.0	
0.4 - 0.6	116.0 ± 3.0 stat ± 5.0 sys	119.0 ± 5.0	115.0 $^{+4.0}_{-8.0}$	98.0 $^{+3.0}_{-5.0}$	116.0 ± 5.0	
0.6 - 0.8	123.0 ± 3.0 stat ± 5.0 sys	126.0 ± 5.0	122.0 $^{+4.0}_{-8.0}$	103.0 $^{+3.0}_{-5.0}$	123.0 ± 5.0	
0.8 - 1.0	133.0 ± 3.0 stat ± 5.0 sys	138.0 $^{+5.0}_{-6.0}$	132.0 $^{+4.0}_{-8.0}$	115.0 $^{+4.0}_{-3.0}$	134.0 ± 5.0	
1.0 - 1.2	136.0 ± 3.0 stat ± 6.0 sys	146.0 ± 6.0	140.0 $^{+5.0}_{-8.0}$	128.0 $^{+4.0}_{-3.0}$	145.0 ± 5.0	
1.2 - 1.4	156.0 ± 3.0 stat ± 6.0 sys	164.0 $^{+6.0}_{-7.0}$	153.0 $^{+5.0}_{-7.0}$	144.0 ± 5.0	158.0 ± 5.0	
1.6 - 1.8	166.0 ± 3.0 stat ± 10.0 sys	195.0 ± 9.0	181.0 ± 5.0	179.0 ± 5.0	190.0 ± 4.0	
1.8 - 2.0	197.0 ± 3.0 stat ± 9.0 sys	207.0 $^{+8.0}_{-10.0}$	196.0 $^{+4.0}_{-3.0}$	200.0 $^{+6.0}_{-3.0}$	206.0 ± 4.0	
2.0 - 2.2	224.0 ± 3.0 stat ± 11.0 sys	224.0 $^{+8.0}_{-11.0}$	211.0 $^{+5.0}_{-3.0}$	213.0 $^{+6.0}_{-3.0}$	219.0 ± 4.0	
2.2 - 2.4	210.0 ± 4.0 stat ± 13.0 sys	241.0 $^{+8.0}_{-12.0}$	225.0 $^{+9.0}_{-8.0}$	231.0 $^{+6.0}_{-3.0}$	231.0 ± 5.0	



“Select” di controllo sul nuovo database **MineHep**

-- TEST: prendo tutte le intestazioni delle colonne delle Y della tabella 2

```
select distinct Header
from YAxisFacts yaf
inner join PaperDim ppd on ppd.ID = yaf.PAPER_ID
where InspireId = 1118047 and TableNumber = 2;
```

-- RESULT: KO mi aspetto che sia vuoto l'Header della Y. Ma non lo è



PT(E)		> 35 GEV
RE		P P --> W+- < E+- N UE > X
SQRT(S)		7000.0 GeV
$ \eta $	$ \eta _{-1}$	
0.0 - 0.2	0.0 - 0.2	23.7
0.2 - 0.4	0.0 - 0.2	2.6
0.4 - 0.6	0.0 - 0.2	2.2
0.6 - 0.8	0.0 - 0.2	2.5
0.8 - 1.0	0.0 - 0.2	2.7
1.0 - 1.2	0.0 - 0.2	2.9
1.2 - 1.4	0.0 - 0.2	2.9
1.6 - 1.8	0.0 - 0.2	2.9
1.8 - 2.0	0.0 - 0.2	2.8



Step futuri

Activity	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
Reverse engineering HepData												
Data population and retrieval												
Structure query interface												
Investigation into medical fields												

- Sviluppo della dashboard, per rappresentare in modo grafico le informazioni di un database organizzato in uno schema a stella.



- Possibili applicazioni al campo della medicina:

