Troubles with Strange Data Structures and Database Growth

Francesco Garue, APL Italiana

ALM: a brief introduction

- Asset-Liability Matching: forecast of assets and liabilities evolution to show that the insurance company’s investments are able
  - To guarantee a target rate of return
  - To meet, in terms of amount and deadlines, the payments due to the policyholders
- Forecast length: at least 15 years requested
- Time unit: 1 month

ALM: a brief introduction

- Assets: computation performed on single holdings
  - Holdings per portfolio range from 100 to 3000
- Liabilities: computation performed on "model-points"
  - Policies per portfolio up to 1M in some cases
  - The clustering process has un upper limit of 12000 model-points

The Audit-Mania

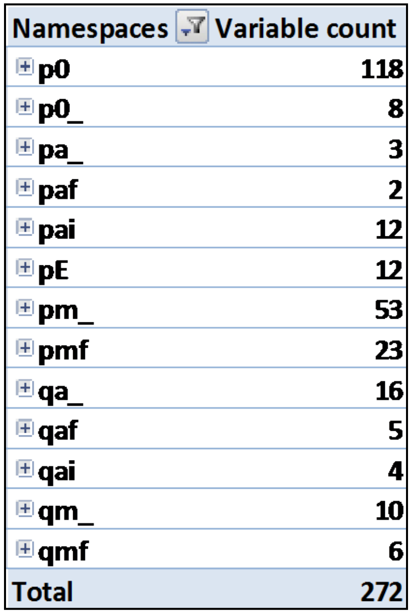
- Increasing demand for insurance companies to be able to trace every internal process: this means saving a lot of stuff to explain how we get from the input data to the output results
- Before 2013 – saving values for each field and month only for the whole portfolio and for some specific model-point classes → data inside the workspace
- Since 2013 – saving the same values for every single model-point: jumping from 10-15 values for each field and month to 10000-12000 → data outside the workspace

The Data Structure

- One big container of all functions and parameters: namespace Alm
- Variables filled during the computation are organized in a set of child namespaces

p0	pai	Pai	pmi	Pmi
p0_	pa_	Pa_	pm_	Pm_
pE	paf	Paf	pmf	Pmf

- Variables are vectors or 3-rows matrices, named with a “short” description of their content.



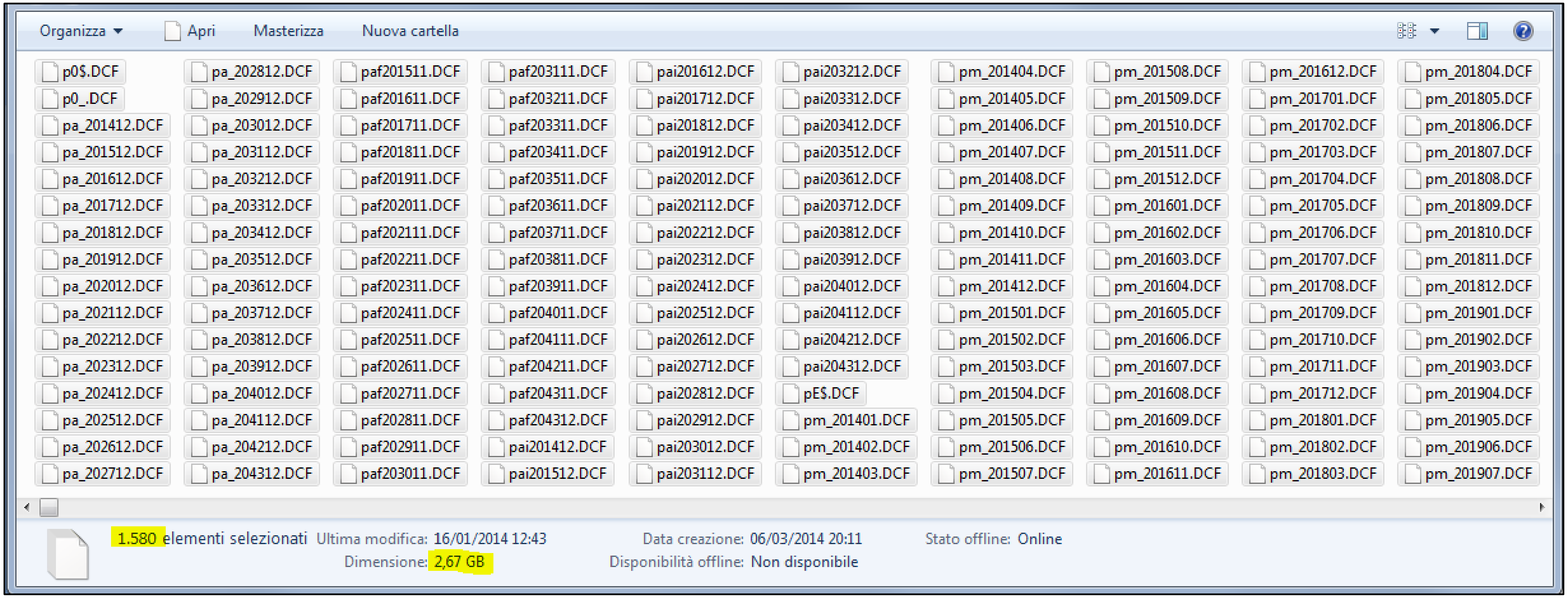
Namespaces	Variable count
⊞ p0	118
⊞ p0_	8
⊞ pa_	3
⊞ paf	2
⊞ pai	12
⊞ pE	12
⊞ pm_	53
⊞ pmf	23
⊞ qa_	16
⊞ qaf	5
⊞ qai	4
⊞ qm_	10
⊞ qmf	6
Total	272

The Storage Files

- Original approach: saving each namespace in the first component of a DCF file
- File names were a short description of their content:
  - The first 3 chars were the namespace name
    - “p0” and “pE” extended with a “$” char
    - “P”s became “q”s
  - The last 6 chars were the year and month of successive savings of the same namespace

The Storage Files

- Can you guess how many files for a 30-year-long forecast?

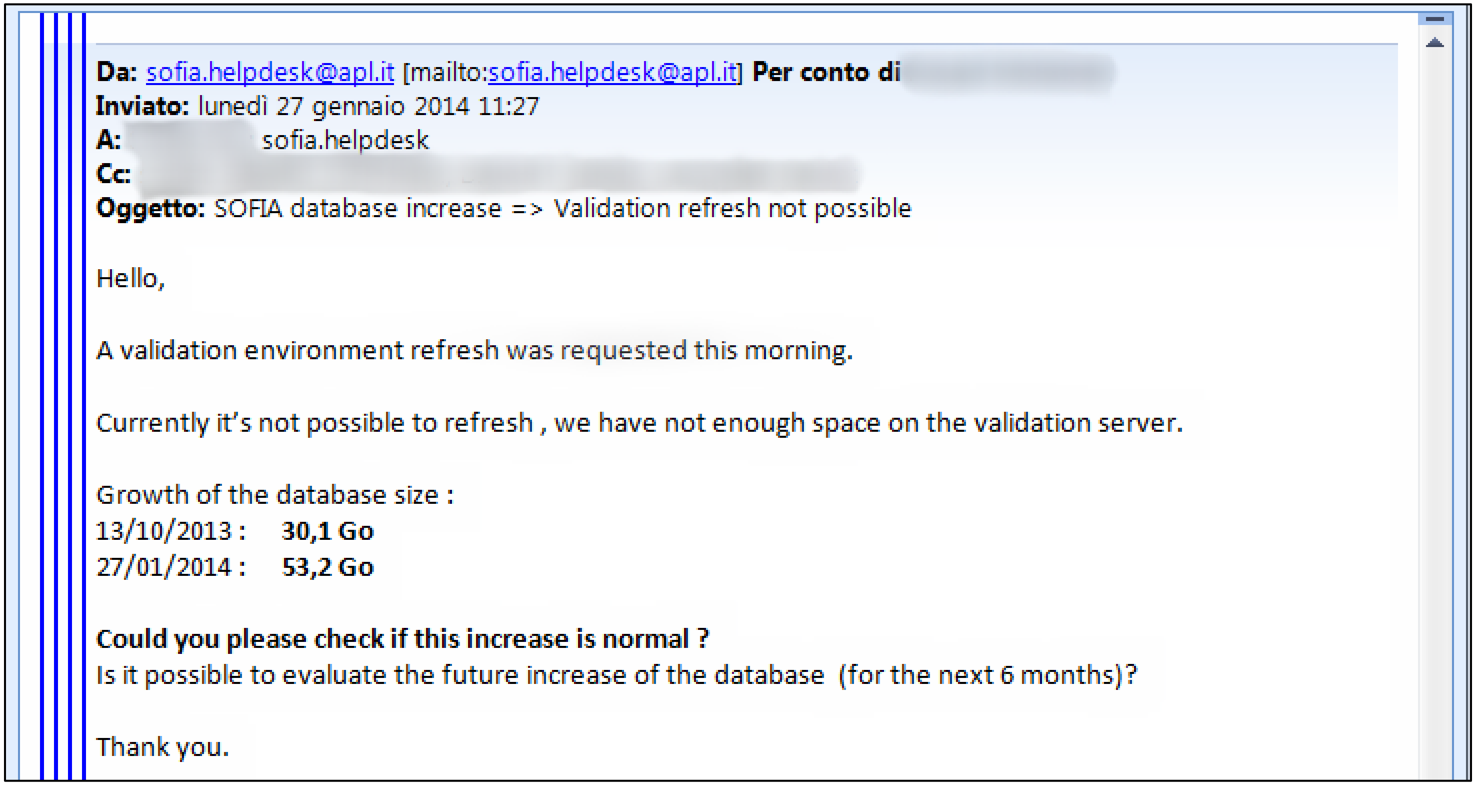


```
p0$.DCF
p0_.DCF
pa_201412.DCF
pa_201512.DCF
pa_201612.DCF
pa_201712.DCF
pa_201812.DCF
pa_201912.DCF
pa_202012.DCF
pa_202112.DCF
pa_202212.DCF
pa_202312.DCF
pa_202412.DCF
pa_202512.DCF
pa_202612.DCF
pa_202712.DCF
pa_202812.DCF
pa_202912.DCF
pa_203012.DCF
pa_203112.DCF
pa_203212.DCF
pa_203312.DCF
pa_203412.DCF
pa_203512.DCF
pa_203612.DCF
pa_203712.DCF
pa_203812.DCF
pa_203912.DCF
pa_204012.DCF
pa_204112.DCF
pa_204212.DCF
pa_204312.DCF
paf201511.DCF
paf201611.DCF
paf201711.DCF
paf201811.DCF
paf201911.DCF
paf202011.DCF
paf202111.DCF
paf202211.DCF
paf202311.DCF
paf202411.DCF
paf202511.DCF
paf202611.DCF
paf202711.DCF
paf202811.DCF
paf202911.DCF
paf203011.DCF
paf203111.DCF
paf203211.DCF
paf203311.DCF
paf203411.DCF
paf203511.DCF
paf203611.DCF
paf203711.DCF
paf203811.DCF
paf203911.DCF
paf204011.DCF
paf204111.DCF
paf204211.DCF
paf204311.DCF
paf204312.DCF
pai201412.DCF
pai201512.DCF
pai201612.DCF
pai201712.DCF
pai201812.DCF
pai201912.DCF
pai202012.DCF
pai202112.DCF
pai202212.DCF
pai202312.DCF
pai202412.DCF
pai202512.DCF
pai202612.DCF
pai202712.DCF
pai202812.DCF
pai202912.DCF
pai203012.DCF
pai203112.DCF
pai203212.DCF
pai203312.DCF
pai203412.DCF
pai203512.DCF
pai203612.DCF
pai203712.DCF
pai203812.DCF
pai203912.DCF
pai204012.DCF
pai204112.DCF
pai204212.DCF
pai204312.DCF
pE$.DCF
pm_201401.DCF
pm_201402.DCF
pm_201403.DCF
pm_201404.DCF
pm_201405.DCF
pm_201406.DCF
pm_201407.DCF
pm_201408.DCF
pm_201409.DCF
pm_201410.DCF
pm_201411.DCF
pm_201412.DCF
pm_201501.DCF
pm_201502.DCF
pm_201503.DCF
pm_201504.DCF
pm_201505.DCF
pm_201506.DCF
pm_201507.DCF
pm_201508.DCF
pm_201509.DCF
pm_201510.DCF
pm_201511.DCF
pm_201512.DCF
pm_201601.DCF
pm_201602.DCF
pm_201603.DCF
pm_201604.DCF
pm_201605.DCF
pm_201606.DCF
pm_201607.DCF
pm_201608.DCF
pm_201609.DCF
pm_201610.DCF
pm_201611.DCF
pm_201612.DCF
pm_201701.DCF
pm_201702.DCF
pm_201703.DCF
pm_201704.DCF
pm_201705.DCF
pm_201706.DCF
pm_201707.DCF
pm_201708.DCF
pm_201709.DCF
pm_201710.DCF
pm_201711.DCF
pm_201712.DCF
pm_201801.DCF
pm_201802.DCF
pm_201803.DCF
pm_201804.DCF
pm_201805.DCF
pm_201806.DCF
pm_201807.DCF
pm_201808.DCF
pm_201809.DCF
pm_201810.DCF
pm_201811.DCF
pm_201812.DCF
pm_201901.DCF
pm_201902.DCF
pm_201903.DCF
pm_201904.DCF
pm_201905.DCF
pm_201906.DCF
pm_201907.DCF
```

1.580 elementi selezionati Ultima modifica: 16/01/2014 12:43 Data creazione: 06/03/2014 20:11 Stato offline: Online
Dimensione: 2,67 GB Disponibilità offline: Non disponibile

BOOM!

- 1st symptom: unexplainable database growth



Da: sofia.helpdesk@apl.it [mailto:sofia.helpdesk@apl.it] **Per conto di** [redacted]
Inviato: lunedì 27 gennaio 2014 11:27
A: [redacted] sofia.helpdesk
Cc: [redacted]
Oggetto: SOFIA database increase => Validation refresh not possible

Hello,

A validation environment refresh was requested this morning.

Currently it's not possible to refresh , we have not enough space on the validation server.

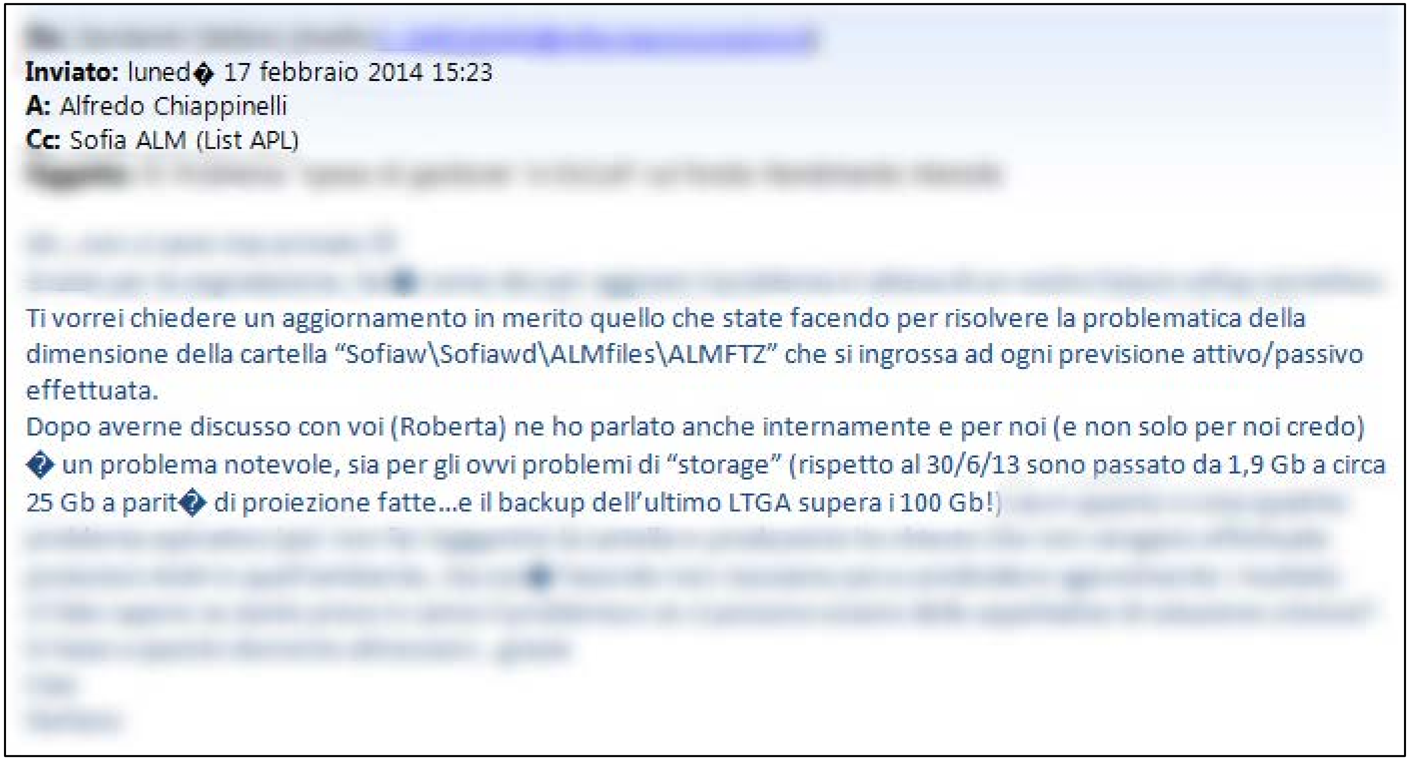
Growth of the database size :
13/10/2013 : **30,1 Go**
27/01/2014 : **53,2 Go**

Could you please check if this increase is normal ?
Is it possible to evaluate the future increase of the database (for the next 6 months)?

Thank you.

BOOM!

- 2nd symptom: explainable database growth

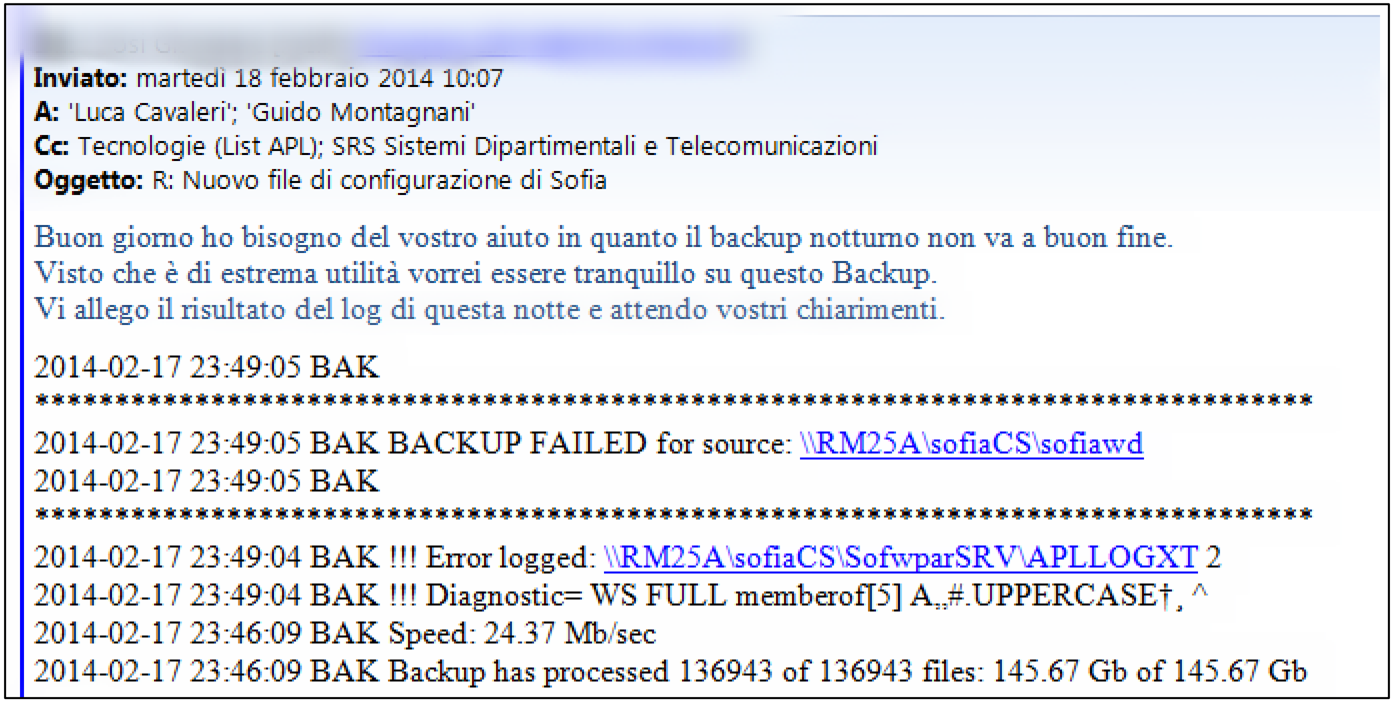


[redacted]
Inviato: luned� 17 febbraio 2014 15:23
A: Alfredo Chiappinelli
Cc: Sofia ALM (List APL)
[redacted]

[redacted]
Ti vorrei chiedere un aggiornamento in merito quello che state facendo per risolvere la problematica della dimensione della cartella "Sofiaw\Sofiawd\ALMfiles\ALMFTZ" che si ingrossa ad ogni previsione attivo/passivo effettuata.
Dopo averne discusso con voi (Roberta) ne ho parlato anche internamente e per noi (e non solo per noi credo) � un problema notevole, sia per gli ovvi problemi di "storage" (rispetto al 30/6/13 sono passato da 1,9 Gb a circa 25 Gb a parit� di proiezione fatte…e il backup dell'ultimo LTGA supera i 100 Gb!) [redacted]

BOOM!

- 3rd symptom: not enough memory



[redacted]

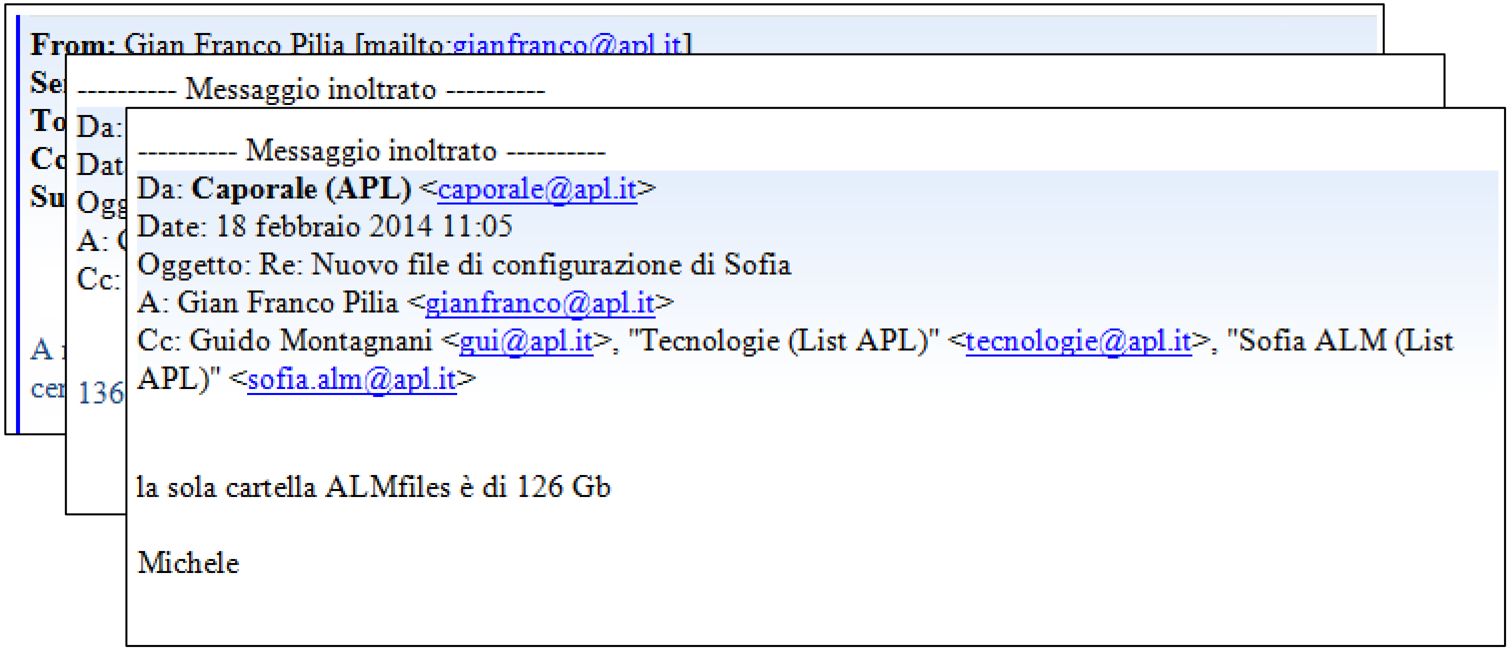
Inviato: martedì 18 febbraio 2014 10:07
A: 'Luca Cavaleri'; 'Guido Montagnani'
Cc: Tecnologie (List APL); SRS Sistemi Dipartimentali e Telecomunicazioni
Oggetto: R: Nuovo file di configurazione di Sofia

Buon giorno ho bisogno del vostro aiuto in quanto il backup notturno non va a buon fine.
Visto che è di estrema utilità vorrei essere tranquillo su questo Backup.
Vi allego il risultato del log di questa notte e attendo vostri chiarimenti.

```
2014-02-17 23:49:05 BAK
********************************************************************************
2014-02-17 23:49:05 BAK BACKUP FAILED for source: \\RM25A\sofiaCS\sofiawd
2014-02-17 23:49:05 BAK
********************************************************************************
2014-02-17 23:49:04 BAK !!! Error logged: \\RM25A\sofiaCS\SofwparSRV\APLLOGXT 2
2014-02-17 23:49:04 BAK !!! Diagnostic= WS FULL memberof[5] A„#.UPPERCASE†, ^
2014-02-17 23:46:09 BAK Speed: 24.37 Mb/sec
2014-02-17 23:46:09 BAK Backup has processed 136943 of 136943 files: 145.67 Gb of 145.67 Gb
```

BOOM!

- 1st reaction: «You cannot be serious!»



From: Gian Franco Pilia [mailto:gianfranco@apl.it]
Se
To
Cc
Su

A r
cer

---------- Messaggio inoltrato ----------
Da:
Dat
Ogg
A: (
Cc:

136

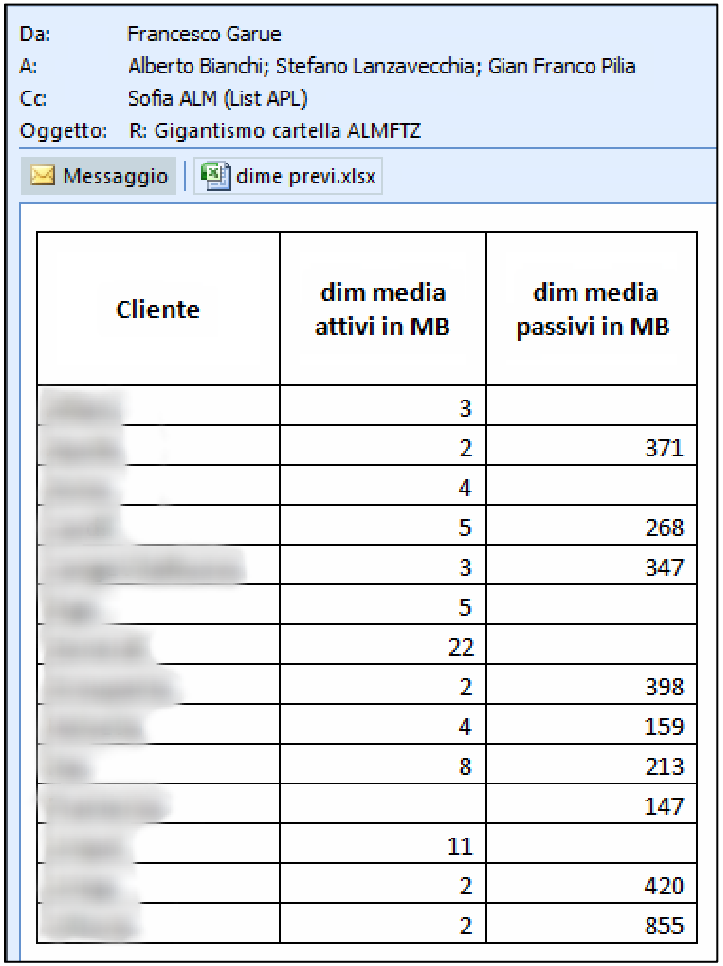
---------- Messaggio inoltrato ----------
Da: **Caporale (APL)** <caporale@apl.it>
Date: 18 febbraio 2014 11:05
Oggetto: Re: Nuovo file di configurazione di Sofia
A: Gian Franco Pilia <gianfranco@apl.it>
Cc: Guido Montagnani <gui@apl.it>, "Tecnologie (List APL)" <tecnologie@apl.it>, "Sofia ALM (List APL)" <sofia.alm@apl.it>

la sola cartella ALMfiles è di 126 Gb

Michele

BOOM!

- 2nd reaction: damage assessment
  - «Are all customers in trouble?» **YES**



Da: Francesco Garue
A: Alberto Bianchi; Stefano Lanzavecchia; Gian Franco Pilia
Cc: Sofia ALM (List APL)
Oggetto: R: Gigantismo cartella ALMFTZ

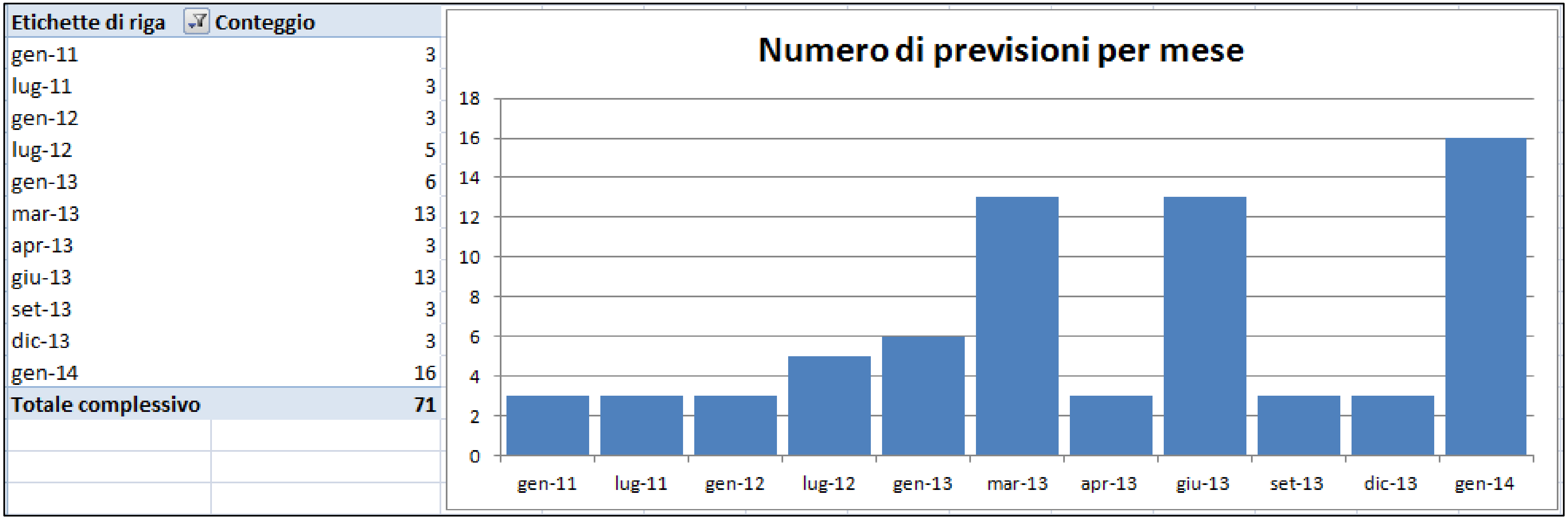
Cliente	dim media attivi in MB	dim media passivi in MB
[redacted]	3	
[redacted]	2	371
[redacted]	4	
[redacted]	5	268
[redacted]	3	347
[redacted]	5	
[redacted]	22	
[redacted]	2	398
[redacted]	4	159
[redacted]	8	213
[redacted]		147
[redacted]	11	
[redacted]	2	420
[redacted]	2	855

BOOM!

- 2nd reaction: damage assessment
  - «Can we estimate the growth rate?» **MAYBE**
    - It depends
      - on the number of portfolios
      - on the model-point features of each portfolio
      - on how (much) the customer uses the software
    - The estimate is acceptable in the near future as long as these parameters don't change

BOOM!

- 2nd reaction: damage assessment
  - «Can we estimate the growth rate?» **MAYBE**

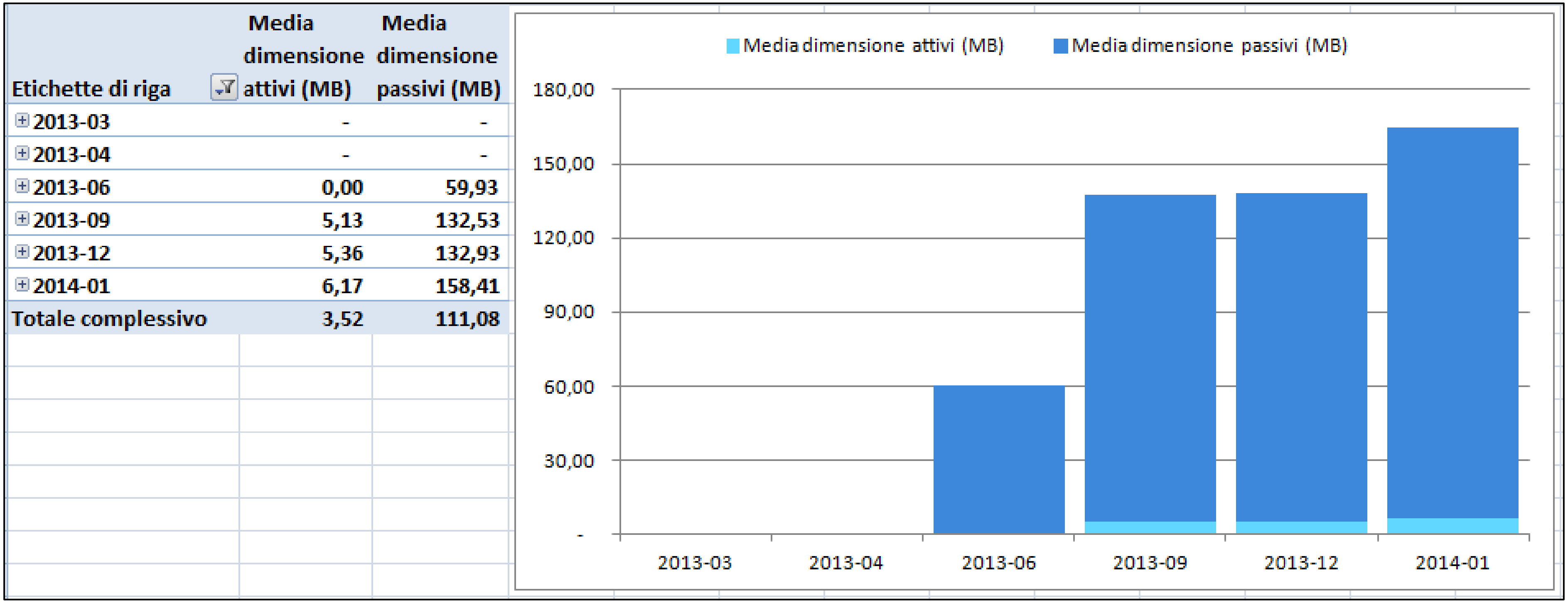


Etichette di riga	Conteggio
gen-11	3
lug-11	3
gen-12	3
lug-12	5
gen-13	6
mar-13	13
apr-13	3
giu-13	13
set-13	3
dic-13	3
gen-14	16
Totale complessivo	**71**

Numero di previsioni per mese

BOOM!

- 2nd reaction: damage assessment
  - «Can we estimate the growth rate?» **MAYBE**



Etichette di riga	Media dimensione attivi (MB)	Media dimensione passivi (MB)
⊞ 2013-03	-	-
⊞ 2013-04	-	-
⊞ 2013-06	0,00	59,93
⊞ 2013-09	5,13	132,53
⊞ 2013-12	5,36	132,93
⊞ 2014-01	6,17	158,41
Totale complessivo	3,52	111,08

Solutions?

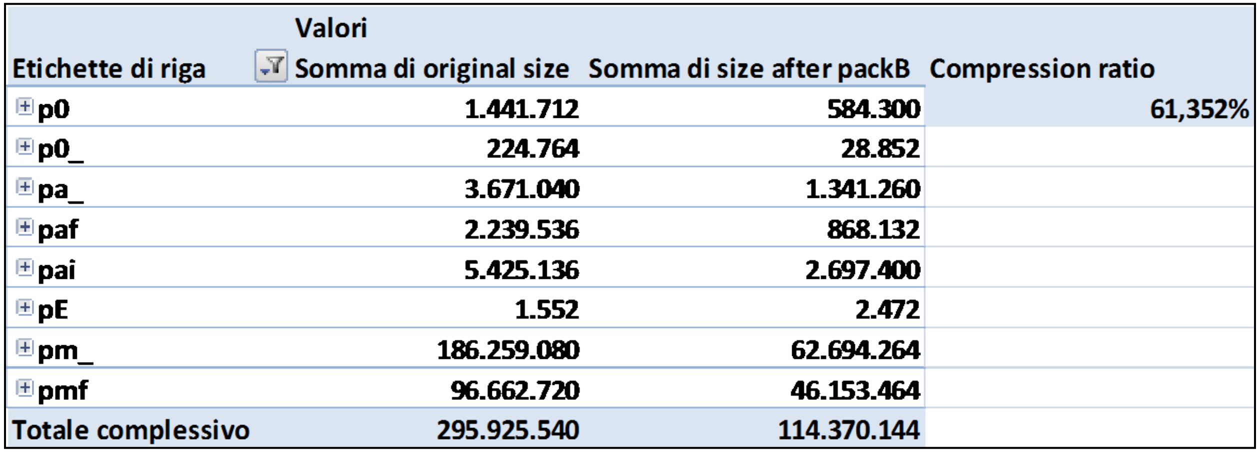
- Choose whether to save all that stuff or not

Solutions?

- Activate Windows file compression for the storage folder
  - Simple, quick solution
  - Read and write not significantly slowed down
  - Experiments showed a compression ratio up to 50%-60% → troubles would have come back again in a few months

Solutions?

- Use packB on each variable
  - Still a quite simple solution
  - Read and write slowed down
  - Experiments showed a compression ratio up to 60% → still not a (good) solution

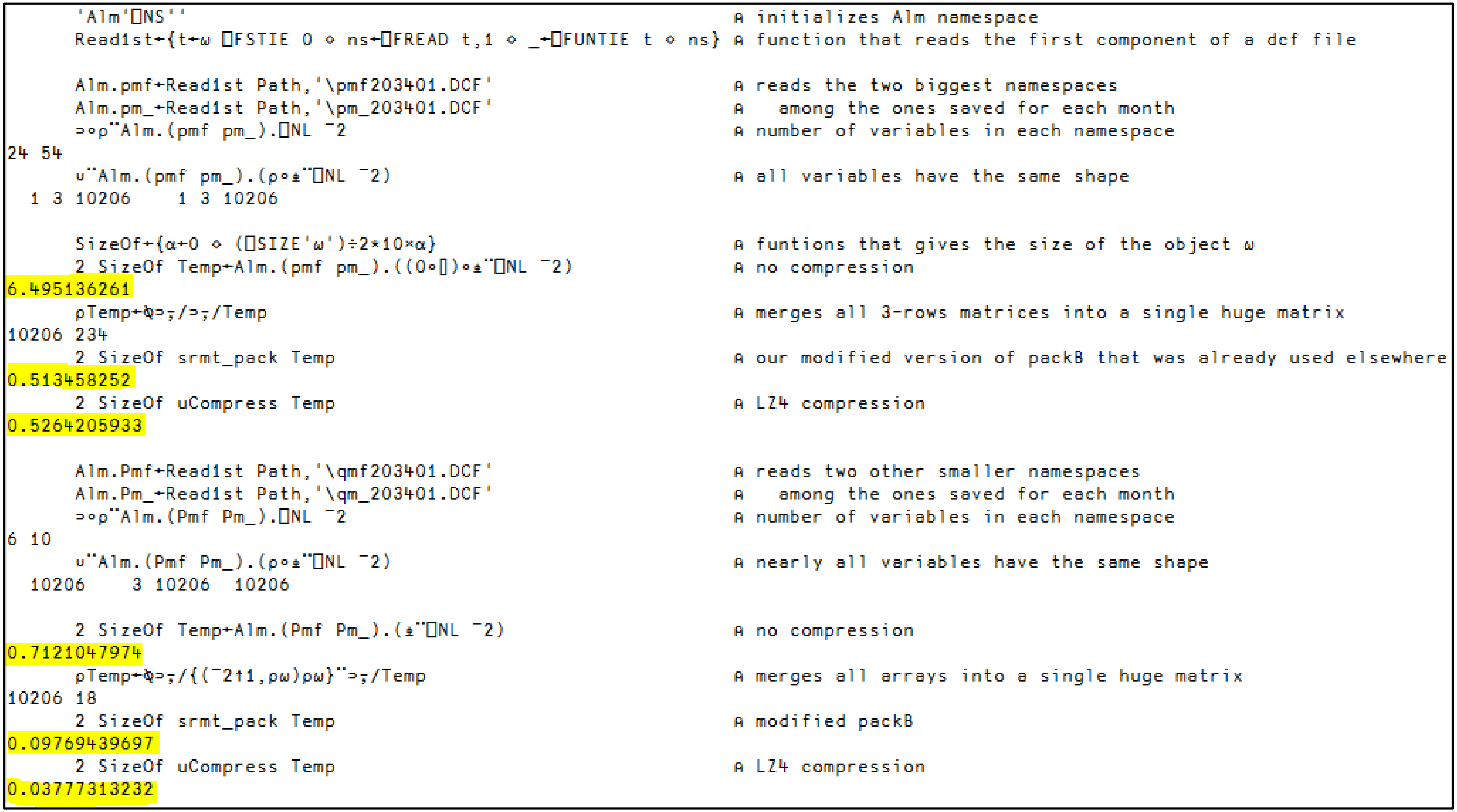


	Valori		
Etichette di riga	Somma di original size	Somma di size after packB	Compression ratio
⊞ p0	1.441.712	584.300	61,352%
⊞ p0_	224.764	28.852	
⊞ pa_	3.671.040	1.341.260	
⊞ paf	2.239.536	868.132	
⊞ pai	5.425.136	2.697.400	
⊞ pE	1.552	2.472	
⊞ pm_	186.259.080	62.694.264	
⊞ pmf	96.662.720	46.153.464	
Totale complessivo	295.925.540	114.370.144	

Solutions?

- Experiments with a few big matrices instead of many small variables (which have one common dimension)
  - Comparing LZ4 with our modified version of packB
  - Results were related to:
    - The model-point features
    - The length of the forecast

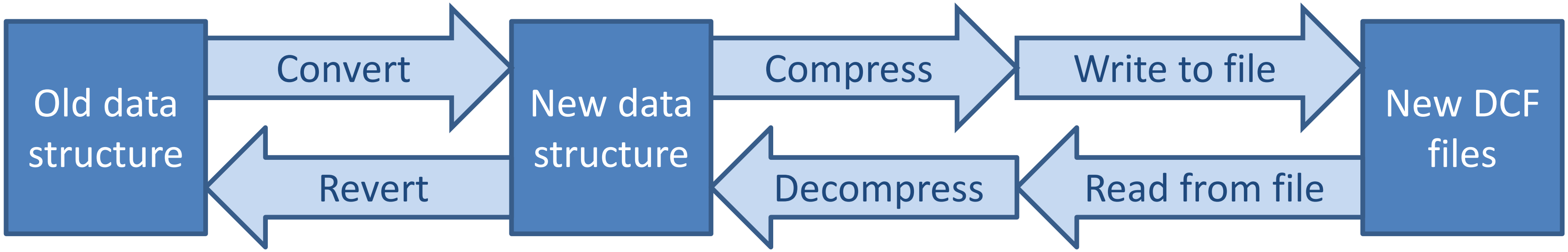
Solutions?



```
      'Alm'⎕NS''                                                ⍝ initializes Alm namespace
      Read1st←{t←⍵ ⎕FSTIE 0 ⋄ ns←⎕FREAD t,1 ⋄ _←⎕FUNTIE t ⋄ ns} ⍝ function that reads the first component of a dcf file

      Alm.pmf←Read1st Path,'\pmf203401.DCF'                     ⍝ reads the two biggest namespaces
      Alm.pm_←Read1st Path,'\pm_203401.DCF'                     ⍝   among the ones saved for each month
      ⊃∘⍴¨Alm.(pmf pm_).⎕NL ¯2                                  ⍝ number of variables in each namespace
24 54
      ∪¨Alm.(pmf pm_).(⍴∘⍎¨⎕NL ¯2)                              ⍝ all variables have the same shape
  1 3 10206    1 3 10206

      SizeOf←{⍺←0 ⋄ (⎕SIZE'⍵')÷2*10×⍺}                         ⍝ funtions that gives the size of the object ⍵
      2 SizeOf Temp←Alm.(pmf pm_).((0∘⌷)∘⍎¨⎕NL ¯2)              ⍝ no compression
6.495136261
      ⍴Temp←⍉⊃⍪/⊃⍪/Temp                                         ⍝ merges all 3-rows matrices into a single huge matrix
10206 234
      2 SizeOf srmt_pack Temp                                   ⍝ our modified version of packB that was already used elsewhere
0.513458252
      2 SizeOf uCompress Temp                                   ⍝ LZ4 compression
0.5264205933

      Alm.Pmf←Read1st Path,'\qmf203401.DCF'                     ⍝ reads two other smaller namespaces
      Alm.Pm_←Read1st Path,'\qm_203401.DCF'                     ⍝   among the ones saved for each month
      ⊃∘⍴¨Alm.(Pmf Pm_).⎕NL ¯2                                  ⍝ number of variables in each namespace
6 10
      ∪¨Alm.(Pmf Pm_).(⍴∘⍎¨⎕NL ¯2)                              ⍝ nearly all variables have the same shape
  10206    3 10206  10206

      2 SizeOf Temp←Alm.(Pmf Pm_).(⍎¨⎕NL ¯2)                    ⍝ no compression
0.7121047974
      ⍴Temp←⍉⊃⍪/{(¯2↑1,⍴⍵)⍴⍵}¨⊃⍪/Temp                           ⍝ merges all arrays into a single huge matrix
10206 18
      2 SizeOf srmt_pack Temp                                   ⍝ modified packB
0.09769439697
      2 SizeOf uCompress Temp                                   ⍝ LZ4 compression
0.03777313232
```

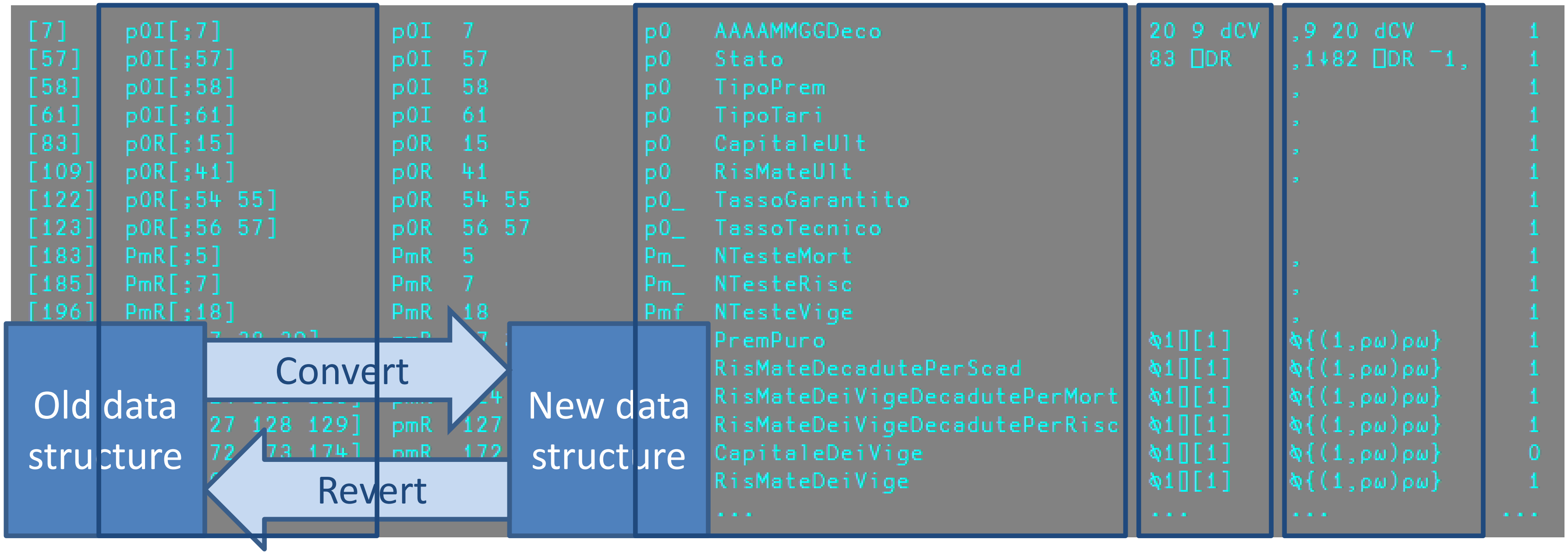
The Resolution

- Given the experiments listed before, changing the data structure from many small variables to a few big matrices seemed necessary
- Considering both the urgency and the extent of the code involved, we decided to convert the data structure before writing to and after reading from file



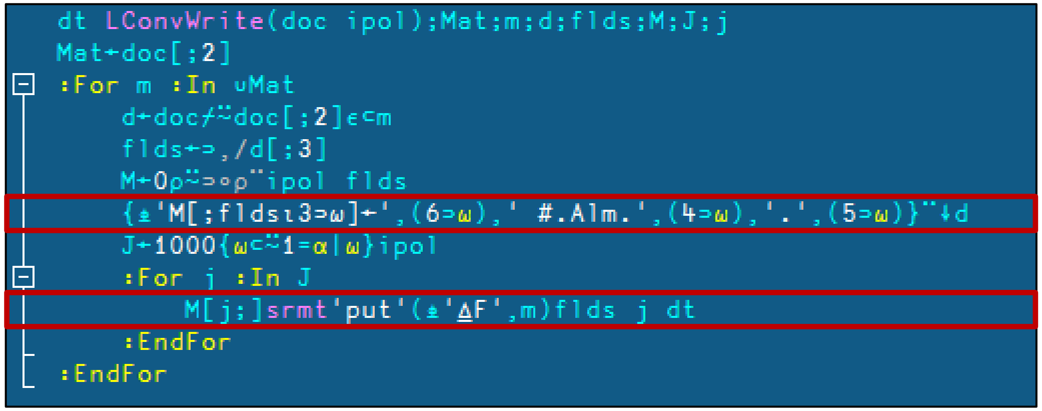
The Resolution

- The Rosetta Stone

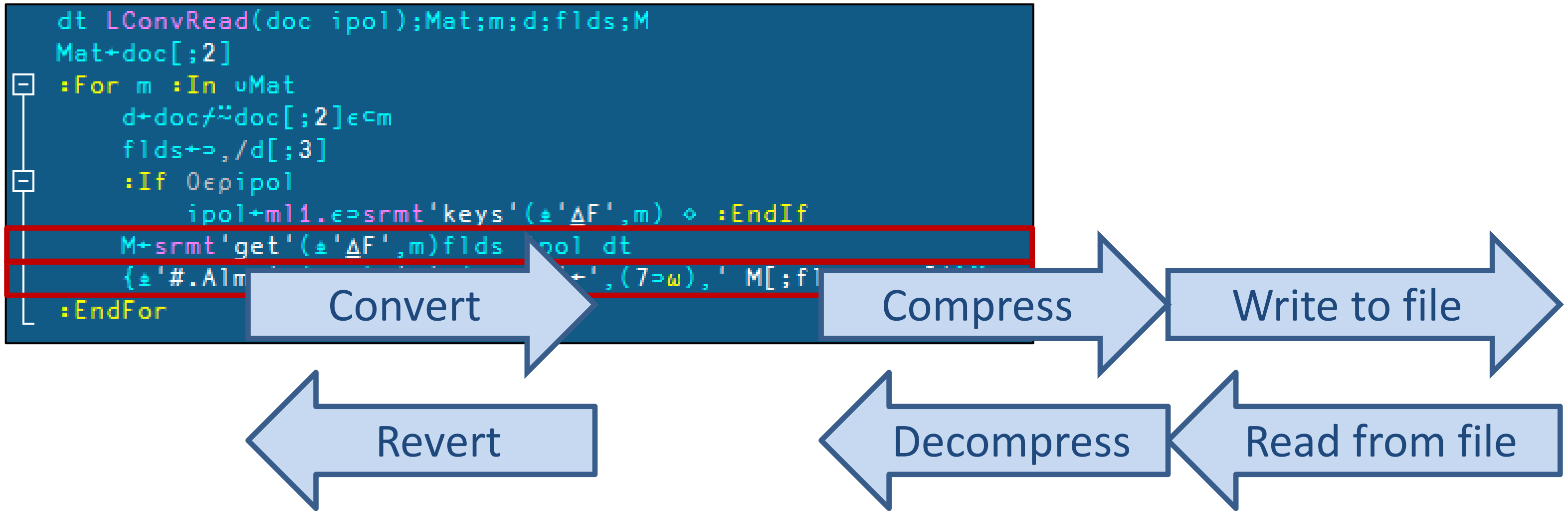


```
[7]     p0I[;7]          p0I  7          p0   AAAAMMGGDeco                      20 9 dCV   ,9 20 dCV       1
[57]    p0I[;57]         p0I  57         p0   Stato                             83 ⎕DR     ,1↓82 ⎕DR ¯1,   1
[58]    p0I[;58]         p0I  58         p0   TipoPrem                                     ,               1
[61]    p0I[;61]         p0I  61         p0   TipoTari                                     ,               1
[83]    p0R[;15]         p0R  15         p0   CapitaleUlt                                  ,               1
[109]   p0R[;41]         p0R  41         p0   RisMateUlt                                   ,               1
[122]   p0R[;54 55]      p0R  54 55      p0_  TassoGarantito                                               1
[123]   p0R[;56 57]      p0R  56 57      p0_  TassoTecnico                                                 1
[183]   PmR[;5]          PmR  5          Pm_  NTesteMort                                   ,               1
[185]   PmR[;7]          PmR  7          Pm_  NTesteRisc                                   ,               1
[196]   PmR[;18]         PmR  18         Pmf  NTesteVige                                   ,               1
                                              PremPuro                          ⍉1⌷[1]     ⍉{(1,⍴⍵)⍴⍵}     1
                                              RisMateDecadutePerScad            ⍉1⌷[1]     ⍉{(1,⍴⍵)⍴⍵}     1
                                              RisMateDeiVigeDecadutePerMort     ⍉1⌷[1]     ⍉{(1,⍴⍵)⍴⍵}     1
           27 128 129]   pmR  127             RisMateDeiVigeDecadutePerRisc     ⍉1⌷[1]     ⍉{(1,⍴⍵)⍴⍵}     1
                                              CapitaleDeiVige                   ⍉1⌷[1]     ⍉{(1,⍴⍵)⍴⍵}     0
                                              RisMateDeiVige                    ⍉1⌷[1]     ⍉{(1,⍴⍵)⍴⍵}     1
                                              ...                               ...        ...           ...
```

The Resolution



```
 dt LConvWrite(doc ipol);Mat;m;d;flds;M;J;j
 Mat←doc[;2]
 :For m :In ∪Mat
     d←doc⌿⍨doc[;2]∊⊂m
     flds←⊃,/d[;3]
     M←0⍴⍨⊃∘⍴¨ipol flds
     {⍎'M[;flds⍳3⊃⍵]←',(6⊃⍵),' #.Alm.',(4⊃⍵),'.',(5⊃⍵)}¨↓d
     J←1000{⍵⊂⍨1=⍺|⍵}ipol
     :For j :In J
         M[j;]srmt'put'(⍎'∆F',m)flds j dt
     :EndFor
 :EndFor
```

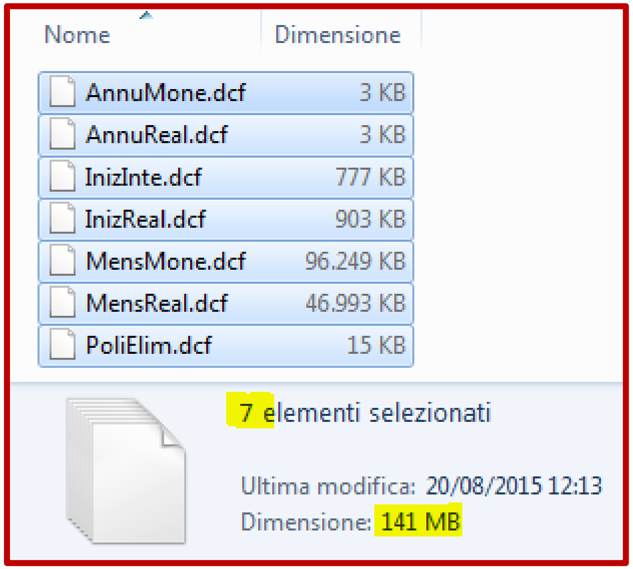


```
 dt LConvRead(doc ipol);Mat;m;d;flds;M
 Mat←doc[;2]
 :For m :In ∪Mat
     d←doc⌿⍨doc[;2]∊⊂m
     flds←⊃,/d[;3]
     :If 0∊⍴ipol
         ipol←ml1.∊⊃srmt'keys'(⍎'∆F',m) ⋄ :EndIf
     M←srmt'get'(⍎'∆F',m)flds ipol dt
     {⍎'#.Alm ←',(7⊃⍵),' M[;fl
 :EndFor
```

Results

- Example of single forecast compression

Data structure	Comp. type	Size (MB)	File count	Comp. ratio
Old	None	2741,44	1580	0%
New	Modified packB	141,54	7	94,837%
New	LZ4	131,99	7	95,185%

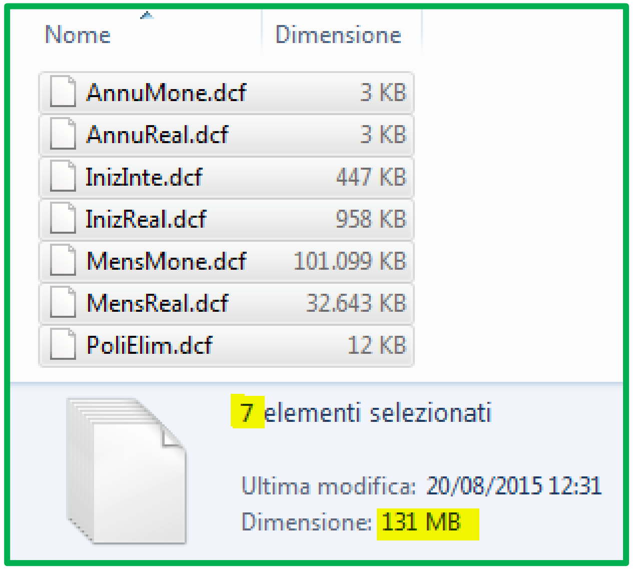


Nome	Dimensione
AnnuMone.dcf	3 KB
AnnuReal.dcf	3 KB
InizInte.dcf	777 KB
InizReal.dcf	903 KB
MensMone.dcf	96.249 KB
MensReal.dcf	46.993 KB
PoliElim.dcf	15 KB

7 elementi selezionati

Ultima modifica: 20/08/2015 12:13

Dimensione: 141 MB



Nome	Dimensione
AnnuMone.dcf	3 KB
AnnuReal.dcf	3 KB
InizInte.dcf	447 KB
InizReal.dcf	958 KB
MensMone.dcf	101.099 KB
MensReal.dcf	32.643 KB
PoliElim.dcf	12 KB

7 elementi selezionati

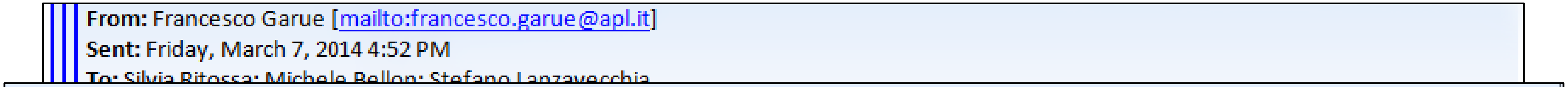
Ultima modifica: 20/08/2015 12:31

Dimensione: 131 MB

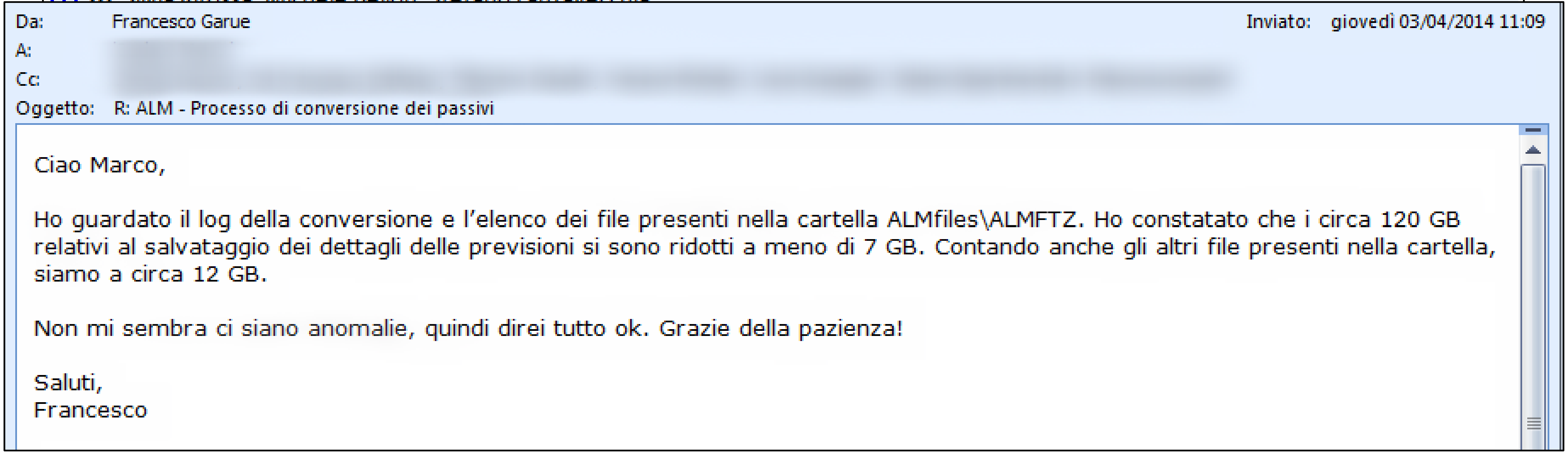
Results

- Example of overall database compression

Data structure	Comp. type	Size (GB)	File count	Comp. ratio
Old	None	119,37	126178	0%
New	Modified packB	5,95	856	95,015%



From: Francesco Garue [mailto:francesco.garue@apl.it]
Sent: Friday, March 7, 2014 4:52 PM
To: Silvia Ritossa; Michele Bellon; Stefano Lanzavecchia



Da: Francesco Garue Inviato: giovedì 03/04/2014 11:09
A: [redacted]
Cc: [redacted]
Oggetto: R: ALM - Processo di conversione dei passivi

Ciao Marco,

Ho guardato il log della conversione e l'elenco dei file presenti nella cartella ALMfiles\ALMFTZ. Ho constatato che i circa 120 GB relativi al salvataggio dei dettagli delle previsioni si sono ridotti a meno di 7 GB. Contando anche gli altri file presenti nella cartella, siamo a circa 12 GB.

Non mi sembra ci siano anomalie, quindi direi tutto ok. Grazie della pazienza!

Saluti,
Francesco