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 an 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

<table>
<thead>
<tr>
<th>p0</th>
<th>pai</th>
<th>Pai</th>
<th>pmi</th>
<th>Pmi</th>
</tr>
</thead>
<tbody>
<tr>
<td>p0_</td>
<td>pa_</td>
<td>Pa_</td>
<td>pm_</td>
<td>Pm_</td>
</tr>
<tr>
<td>pE</td>
<td>paf</td>
<td>Paf</td>
<td>pmf</td>
<td>Pmf</td>
</tr>
</tbody>
</table>

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

- 1\textsuperscript{st} symptom: unexplainable database growth

---

**Email:**

**Da:** sofia.helpdesk@apl.it [mailto:sofia.helpdesk@apl.it] Per conto di

**Inviato:** luned\'i 27 gennaio 2014 11:27

**A:** sofia.helpdesk

**Cc:**

**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.
• 2<sup>nd</sup> symptom: explainable database growth
BOOM!

- 3rd symptom: not enough memory
BOOM!

• 1st reaction: «You cannot be serious!»
BOOM!

- 2\textsuperscript{nd} reaction: damage assessment
  - «Are all customers in trouble?» YES
• 2\textsuperscript{nd} 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
BOOM!

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

<table>
<thead>
<tr>
<th>Etichette di riga</th>
<th>Somma di original size</th>
<th>Somma di size after packB</th>
<th>Compression ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>p0</td>
<td>1.441.712</td>
<td>584.300</td>
<td>61,352%</td>
</tr>
<tr>
<td>p0_</td>
<td>224.764</td>
<td>28.852</td>
<td></td>
</tr>
<tr>
<td>pa</td>
<td>3.671.040</td>
<td>1.341.260</td>
<td></td>
</tr>
<tr>
<td>paf</td>
<td>2.239.536</td>
<td>868.132</td>
<td></td>
</tr>
<tr>
<td>pai</td>
<td>5.425.136</td>
<td>2.697.400</td>
<td></td>
</tr>
<tr>
<td>pE</td>
<td>1.552</td>
<td>2.472</td>
<td></td>
</tr>
<tr>
<td>pm_</td>
<td>186.259.080</td>
<td>62.694.264</td>
<td></td>
</tr>
<tr>
<td>pmf</td>
<td>96.662.720</td>
<td>46.153.464</td>
<td></td>
</tr>
<tr>
<td>Totale complessivo</td>
<td>295.925.540</td>
<td>114.370.144</td>
<td></td>
</tr>
</tbody>
</table>
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?

A initializes Alm namespace
A function that reads the first component of a dcf file
A reads the two biggest namespaces
A among the ones saved for each month
A number of variables in each namespace
A all variables have the same shape
A functions that gives the size of the object \( \omega \)
A no compression
A merges all 3-rows matrices into a single huge matrix
A our modified version of pack8 that was already used elsewhere
A LI4 compression
A reads two other smaller namespaces
A among the ones saved for each month
A number of variables in each namespace
A nearly all variables have the same shape
A no compression
A merges all arrays into a single huge matrix
A modified pack8
A LI4 compression
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
The Resolution

```plaintext
dt LConvWrite(doc ipol); Mat;m;d;flds;M;J;j
Mat=doc[];2]
  :For m :In uMat
d=doc[][];2]εm
d+=doc[;3]
M+=ipol.flds[
(ε' M[;flds3=ω]=',(6=ω), '#.Alm.','(4=ω)',':',(5=ω))'"d
J=1000(ω=1=a)ipol
  :For j :In J
    M[jj]srmt'put'(ε' AF',m)flds j dt
  :EndFor
:EndFor

LConvRead(doc ipol); Mat;m;d;flds;M
Mat=doc[];2]
  :For m :In uMat
d=doc[][];2]εm
d+=doc[;3]
  :If 0lipol
    ipol+=m1.e=srmt'keys'(ε' AF',m) :EndIf
    M=srmt'get'(ε' AF',m)flds ipol dt
    (ε' #.Alm.',(7=ω)',M[fld])
  :EndFor
```

Convert  Compress  Write to file

Revert  Decompress  Read from file
Results

- Example of single forecast compression

<table>
<thead>
<tr>
<th>Data structure</th>
<th>Comp. type</th>
<th>Size (MB)</th>
<th>File count</th>
<th>Comp. ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>None</td>
<td>2741,44</td>
<td>1580</td>
<td>0%</td>
</tr>
<tr>
<td>New</td>
<td>Modified packB</td>
<td>141,54</td>
<td>7</td>
<td>94,837%</td>
</tr>
<tr>
<td>New</td>
<td>LZ4</td>
<td>131,99</td>
<td>7</td>
<td>95,185%</td>
</tr>
</tbody>
</table>
Results

- Example of overall database compression

<table>
<thead>
<tr>
<th>Data structure</th>
<th>Comp. type</th>
<th>Size (GB)</th>
<th>File count</th>
<th>Comp. ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old</td>
<td>None</td>
<td>119,37</td>
<td>126178</td>
<td>0%</td>
</tr>
<tr>
<td>New</td>
<td>Modified packB</td>
<td>5,95</td>
<td>856</td>
<td>95,015%</td>
</tr>
</tbody>
</table>

---

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

Da: Francesco Garue
A: 
Cc: 
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