R as a Query Language?

Hannes Mühleisen
• Observation A: *Nobody* loves SQL

• Observation B: *Everybody* loves R / Python / Matlab

• Observation C: People still do the same things
  • Native plus specialised packages, 
    eg. *data.table / dplyr / NumPy / Pandas / ...* 

\[ \sigma \Pi \Join G \]
• Let’s bring some relational know-how to the party
• Interpret a imperative analysis script as a *declaration of intent*
• Then, apply standard DB query optimisations
  • and possibly use optimised operators as well
• Pushdown of point/range selections, projections
• Detection of common subexpressions
  • Copy-paste “pattern”
  • Caching
• Rescheduling
  • Minimise variable life time
  • Parallelisation of independent flows
• Join ordering
\[
\begin{align*}
d &\leftarrow \text{data.frame}(a=\text{seq}(100), b=\text{runif}(100)) \\
d$b &\leftarrow d$b/2 \\
d &\leftarrow d[d$a == 50,] \\
\end{align*}
\]
• Apply default DB rule-based query optimisations

• Implementation:
  • R on the JVM, including some C translation
  • Uses promises/deferred executions

• Use case: Survey analysis (survey package for R)
  • Complex calculations
  • Large datasets

\[
\overline{age} = \frac{\sum (age \times w)}{\sum w}
\]

\[
SE(\overline{age}) = \sqrt{\frac{4}{80} \sum_{r=1}^{80} (w_r - \overline{age})^2}
\]
svydsnn <- svrepdesign(
    weight = ~pwgtp,
    repweights = 'pwgtp[1-9]',
    scale = 4 / 80,
    rscales = rep(1, 80),
    mse = TRUE,
    data = svydata)

svytot(s <- I(agep %in% 0:4) + I(agep %in% 5:9) + I(agep %in% 10:14) + I(agep %in% 15:19), svydsnn),
svytot(s <- I(agep %in% 20:24) + I(agep %in% 25:34) + I(agep %in% 35:44) + I(agep %in% 45:54), svydsnn),
svytot(s <- I(agep %in% 55:59) + I(agep %in% 60:64) + I(agep %in% 65:74) + I(agep %in% 75:84) + I(agep > 84), svydsnn)
Thank you.
Questions?

http://hannes.muehleisen.org
@hfmuehleisen