Large Instance Processing

Mark Goodhand

Madrid, June 2015

CoreFiling
Precision Reporting

How big is ‘large’

- UK iXBRL document ~ 100 KB
- SEC filing with extension ~ 10 MB
- CRDIV ~ 100 MB ?
- SII ~ 1 GB ??

CoreFiling
Precision Reporting

COREP/FINREP – smaller than expected

- Expected 100s of MB, perhaps gigabytes
- Actually under 500 MB
- Vast majority under 10 MB

CoreFiling
Precision Reporting

COREP/FINREP – “the 1%”

- Biggest document: ~ 350 MB
- > 300 MB per quarter: 150
- Total per quarter: 11 400
- Average size: 64 MB

Estimates for European Banking Supervision by Banco de España

CoreFiling
Precision Reporting

Solvency II

- How big will they be?

CoreFiling
Precision Reporting

Solvency II

- 2 GB ?
- Nobody knows for sure

CoreFiling
Precision Reporting

Solvency II

- Good news: we’re ready
- CRD IV prepared us

CoreFiling
Precision Reporting

Large Instance challenges

- Processing time
- Memory use

CoreFiling
Precision Reporting

How fast is fast enough?

- Upload, process, store
  - Minutes, not hours
  - Processing includes Formula and Table Linkbase
- Query / analyse
  - Seconds or milliseconds

Throw hardware at the problem?

- Server
  - 30+ core Xeon
  - 256 GB RAM
- Laptop
  - Quad core i7
  - 16 GB RAM

CoreFiling
Precision Reporting

Algorithms and data structures

- Millions of facts and evaluations
- Naïve approaches can take days
- Prefer not to use gigabytes of RAM

CoreFiling
Precision Reporting

Just use a database?

- Databases are mature, but not magical
- Disks are slow
- Must ensure validation is faithful to standards
    - SQL vs XPath data types and operators
    - Interval arithmetic

Streaming

- Put contexts & units before facts
- Bounded buffers for each
  - for Eurofiling, buffer size 1
- Specification from xbrl.org
  - PWD 2013-03-06
  - CR end Q2
  - REC end Q3

CoreFiling
Precision Reporting

Streaming

- Advantages
  - Process in a single pass with *fixed memory*
  - Fact events processed asynchronously, in parallel
  - Backwards compatible
  - XBRL 2.1 + XDT validation
  - Streaming facts to database

CoreFiling
Precision Reporting

Streaming

- Disadvantages
  - Fixed memory not always achievable
  - Formula Linkbase requires multiple facts
  - Filing rules require state proportional to size of instance
    - E.g. duplicate context/unit checks

Filing Rules validation

- Don’t worry about duplicate contexts / units
- Focus on semantics
- Still concerned about duplicate facts
  - Especially if inconsistent

Formula

- Arbitrary XPath requires full instance XML
- Most assertions are simpler
  - iaf:numeric-equal(iaf:sum($a), $b)
  - iaf:numeric-equal($a, $b)
- Could be backed by non-XML model
  - Open Information Model Working Group

Further reading

- Streaming Extensions 1.0
  - http://specifications.xbrl.org/work-product-index-streaming-extensions-streaming-extensions-1.0.html
- Notes on the Processing of Large Instances
  - http://www.xbrl.org/WGN/large-instance-processing/WGN-2012-10-31/large-instance-processing-WGN-WGN-2012-10-31.html