RDF pro
Processing Billions of RDF Triples on a Single Machine using Streaming and Sorting

Francesco Corcoglioniti, Marco Rospocher, Marco Amadori, Michele Mostarda
Fondazione Bruno Kessler-IRST
Trento, Italy

http://rdfpro.fbk.eu

SAC2015
Salamanca, 14 April 2015
The problem

Are relevant RDF processing tasks on large datasets practically feasible on a single commodity machine by using streaming and sorting techniques?
The problem
The problem

- perform relevant RDF processing tasks
  - TBox and statistics extraction
  - data filtering
  - data transformation
  - inference materialisation
  - smushing
  - ...

3/22
The problem

- perform relevant RDF processing tasks
  - TBox and statistics extraction
  - data filtering
  - data transformation
  - inference materialisation
  - smushing
  - ...

- on large datasets
  - LOD-sized: billions of triples
  - quads, not just triples
The problem

- perform relevant RDF processing tasks
  - TBox and statistics extraction
  - data filtering
  - data transformation
  - inference materialisation
  - smushing
  - ...

- on large datasets
  - LOD-sized: billions of triples
  - quads, not just triples

- on a single commodity machine
  - no cluster / distributed computing
  - no triplestore or other data index
The problem

- perform relevant RDF processing tasks
  - TBox and statistics extraction
  - data filtering
  - data transformation
  - inference materialisation
  - smushing
  - ...

- on large datasets
  - LOD-sized: billions of triples
  - quads, not just triples

- on a single commodity machine
  - no cluster / distributed computing
  - no triplestore or other data index

- using streaming and sorting
  - data processing primitives managing large amounts of data with constrained resources
Our Contributions

- **RDF_{pro}**: an extensible tool for building RDF processing pipelines based on streaming and sorting

- **Empirical Evaluation** on 4 usage scenarios, positively answering our research question
RDF pro
http://rdfpro.fbk.eu
RDF\textsubscript{pro} at its core: RDF processor

- Based on Streaming:
  - quads from the input stream are processed one at a time
  - multiple passes can be performed
  - may have an internal state / side effects (e.g., writing)
RDF_{pro}: sorting

- offered to processors as a **primitive** to arbitrarily sort selected data during a pass
  - implemented via **external sorting** (unix sort + smart data encoding)
  - **effectively exploits** available hardware resources

- **enables tasks** not feasible with streaming alone:
  - duplicates removal
  - set operations
  - any task that need to group together scattered information
**RDF**<sub>pro</sub>: on-board RDF processors

- move data around
  - @read / @write files
  - @download from / @upload to SPARQL endpoints

- transform data
  - arbitrary data @transform while streaming on triples (via Groovy scripts)
  - @infer the RDFS closure
  - @smush data, merging owl:sameAs URIs into canonical URIs
  - extract @tbox and VOID @stats
  - @unique discards duplicates
RDF\textsubscript{pro}: processor composition

- processors can be derived by (recursively) applying sequential

\[
\text{rdfpro} @P_1 \text{ args}_1 \ldots @P_N \text{ args}_N
\]

and parallel compositions

\[
\text{rdfpro} \{ @P_1 \text{ args}_1, \ldots, @P_N \text{ args}_N \} f
\]
**RDF_{pro}: processor composition**

Example

- read a Turtle+gzip file (*file.ttl.gz*)
- **TBox** and **VOID** statistics are extracted in parallel
- union written to an RDF/XML file (*onto.rdf*)
RDF\textsubscript{pro}: further details

- Offered as:
  - Java command line tool
  - embeddable Java library

- Built using a multi-thread design to fully exploit CPU resources

- Built on top of Sesame RDF library

- Extendable with new processors

- Web-site: http://rdfpro.fbk.eu/

- Code
  - available at: https://github.com/dkmfbk/rdfpro
  - CC0 license
Empirical Evaluation
4 usage scenarios

Commodity machine used in all the scenarios:
Intel Core i7 860 CPU (4 cores, hyper-threading)
16 GB RAM
500 GB 7200 RPM hard disk
Linux 2.6.32
Scenario 1: Dataset Analysis

- **TASK:** provide a qualitative and quantitative characterisation of the contents of an RDF dataset (e.g., extract TBox or compute ABox data statistics)
  - to identify relevant data, pre-processing needs
  - to characterise a dataset for validation / documentation

- **EXPERIMENT:** extract TBox and statistics from a version of Freebase
  - 2014/09/10 dump, 2863 millions of quads (MQ)
  - 2014/07/10 dump, 2623 MQ

and compare it with an older version
Scenario 1: Dataset Analysis

1. extract TBox and
2. compute ABox data Statistics

<table>
<thead>
<tr>
<th>Task</th>
<th>Input</th>
<th>Output</th>
<th>Throughput</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[MQ]</td>
<td>[MB]</td>
<td>[MQ]</td>
<td>[MB]</td>
</tr>
<tr>
<td>1. TBox</td>
<td>2863</td>
<td>28339</td>
<td>0.23</td>
<td>3.01</td>
</tr>
<tr>
<td>2. Statistics</td>
<td>2863</td>
<td>28339</td>
<td>0.13</td>
<td>1.36</td>
</tr>
<tr>
<td>1-2 Aggregated</td>
<td>2863</td>
<td>28339</td>
<td>0.36</td>
<td>4.35</td>
</tr>
<tr>
<td>3. Comparison</td>
<td>5486</td>
<td>55093</td>
<td>260</td>
<td>1894</td>
</tr>
</tbody>
</table>

3. Compare datasets
Scenario 1: Dataset Analysis

1. extract TBox and
2. compute ABox data Statistics

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. TBox</td>
<td>2863</td>
<td>28339</td>
<td>0.23</td>
<td>3.01</td>
<td>0.23</td>
<td>3.01</td>
<td>1.43</td>
</tr>
<tr>
<td>2. Statistics</td>
<td>2863</td>
<td>28339</td>
<td>0.13</td>
<td>1.36</td>
<td>0.34</td>
<td>1.36</td>
<td>8443</td>
</tr>
<tr>
<td>1-2 Aggregated</td>
<td>2863</td>
<td>28339</td>
<td>0.36</td>
<td>4.35</td>
<td>0.34</td>
<td>3.36</td>
<td>8426</td>
</tr>
<tr>
<td>3. Comparison</td>
<td>5486</td>
<td>55093</td>
<td>260</td>
<td>1894</td>
<td>0.42</td>
<td>4.25</td>
<td>12955</td>
</tr>
</tbody>
</table>
Scenario 1: Dataset Analysis

1. Extract TBox and
2. Compute ABox data Statistics
3. Compare datasets
Scenario 1: Dataset Analysis

1. extract TBox and
2. compute ABox data Statistics

3. Compare datasets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. TBox</td>
<td>2863</td>
<td>28339</td>
<td>0.23</td>
<td>3.01</td>
<td>1.43</td>
<td>14.12</td>
<td>2006</td>
</tr>
<tr>
<td>2. Statistics</td>
<td>2863</td>
<td>28339</td>
<td>0.13</td>
<td>1.36</td>
<td>0.34</td>
<td>3.36</td>
<td>8443</td>
</tr>
<tr>
<td>1-2 Aggregated</td>
<td>2863</td>
<td>28339</td>
<td>0.36</td>
<td>4.35</td>
<td>0.34</td>
<td>3.36</td>
<td>8426</td>
</tr>
<tr>
<td>3. Comparison</td>
<td>5486</td>
<td>55093</td>
<td>260</td>
<td>1894</td>
<td>0.42</td>
<td>4.25</td>
<td>12955</td>
</tr>
</tbody>
</table>
Scenario 1: Dataset Analysis
Scenario 2: Dataset Filtering

- TASK: extract a subset of data, by
  1. identifying the entities of interest in the dataset (selection conditions on their URIs, rdf:type or other properties)
  2. extracting selected quads about these entities

- EXPERIMENT: extract from Freebase (2014/07/10, 2863 MQ):
  - entities of interest: musical groups (rdf:type = fb:music.musical_group) that are still active (having no fb:music.artist.active_end triples)
  - properties to extract: group name (rdfs:label), genre (fb:music.artist.genre) and place of origin (fb:music.artist.origin)
Scenario 2: Dataset Filtering

1. Select entities: 2863 MQ, 28339 MB, 0.20 MQ, 0.73 MB, 1.36 MQ/s, 13.4 MB/s, Time: 2111 s
2. Extract quads: 2863 MQ, 28339 MB, 0.42 MQ, 5.17 MB, 1.15 MQ/s, 11.4 MB/s, Time: 2481 s
Scenario 2: Dataset Filtering

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Select entities</td>
<td>2863</td>
<td>28339</td>
<td>0.20</td>
<td>0.73</td>
<td>1.36</td>
<td>13.4</td>
<td>2111</td>
</tr>
<tr>
<td>Extract quads</td>
<td>2863</td>
<td>28339</td>
<td>0.42</td>
<td>5.17</td>
<td>1.15</td>
<td>11.4</td>
<td>2481</td>
</tr>
</tbody>
</table>
Scenario 3: Dataset Merging

- TASK: multiple RDF datasets are integrated and prepared for application consumption
  - comprises tasks such as smushing, inference materialization and data deduplication

- EXPERIMENT: merging of
  - Freebase (2014/07/10, 2863 MQ)
  - GeoNames (2013/08/27, 125 MQ)
  - 4 DBpedia subsets (EN, ES, IT, NL - version 3.9, 406 MQ)
Total: 3394 MQ
Scenario 3: Dataset Merging

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>@transform</td>
<td>3394</td>
<td>33524</td>
<td>3394</td>
<td>36903</td>
<td>0.42</td>
<td>4.12</td>
<td>8137</td>
</tr>
<tr>
<td>@tbox</td>
<td>3394</td>
<td>36903</td>
<td>&lt;1</td>
<td>4</td>
<td>1.28</td>
<td>13.9</td>
<td>2656</td>
</tr>
<tr>
<td>@smush</td>
<td>3394</td>
<td>36903</td>
<td>3424</td>
<td>38823</td>
<td>0.37</td>
<td>3.98</td>
<td>9265</td>
</tr>
<tr>
<td>@infer</td>
<td>3424</td>
<td>38823</td>
<td>5615</td>
<td>51927</td>
<td>0.32</td>
<td>3.66</td>
<td>10612</td>
</tr>
<tr>
<td>@unique</td>
<td>5615</td>
<td>51927</td>
<td>4085</td>
<td>31297</td>
<td>0.33</td>
<td>3.03</td>
<td>17133</td>
</tr>
</tbody>
</table>

1. Aggregated
2. Aggregated
Scenario 3: Dataset Merging

![Diagram of dataset merging process]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>@transform</td>
<td>3394</td>
<td>33524</td>
<td>3394</td>
<td>36903</td>
<td>0.42</td>
<td>4.12</td>
<td>8137</td>
</tr>
<tr>
<td>@tbox</td>
<td>3394</td>
<td>36903</td>
<td>&lt;1</td>
<td>4</td>
<td>1.28</td>
<td>13.9</td>
<td>2656</td>
</tr>
<tr>
<td>@smush</td>
<td>3394</td>
<td>36903</td>
<td>3424</td>
<td>38823</td>
<td>0.37</td>
<td>3.98</td>
<td>9265</td>
</tr>
<tr>
<td>@infer</td>
<td>3424</td>
<td>38823</td>
<td>5615</td>
<td>51927</td>
<td>0.32</td>
<td>3.66</td>
<td>10612</td>
</tr>
<tr>
<td>@unique</td>
<td>5615</td>
<td>51927</td>
<td>4085</td>
<td>31297</td>
<td>0.33</td>
<td>3.03</td>
<td>17133</td>
</tr>
</tbody>
</table>

1. Aggregated
2. Aggregated
Scenario 3: Dataset Merging

<table>
<thead>
<tr>
<th>Step</th>
<th>Input MQ</th>
<th>Input MB</th>
<th>Output MQ</th>
<th>Output MB</th>
<th>Throughput MQ/s</th>
<th>Throughput MB/s</th>
<th>Time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>@transform</td>
<td>3394</td>
<td>33524</td>
<td>3394</td>
<td>36903</td>
<td>0.42</td>
<td>4.12</td>
<td>8137</td>
</tr>
<tr>
<td>@tbox</td>
<td>3394</td>
<td>36903</td>
<td>&lt;1</td>
<td>4</td>
<td>1.28</td>
<td>13.9</td>
<td>2656</td>
</tr>
<tr>
<td>@smush</td>
<td>3394</td>
<td>36903</td>
<td>3424</td>
<td>38823</td>
<td>0.37</td>
<td>3.98</td>
<td>9265</td>
</tr>
<tr>
<td>@infer</td>
<td>3424</td>
<td>38823</td>
<td>5615</td>
<td>51927</td>
<td>0.32</td>
<td>3.66</td>
<td>10612</td>
</tr>
<tr>
<td>@unique</td>
<td>5615</td>
<td>51927</td>
<td>4085</td>
<td>31297</td>
<td>0.33</td>
<td>3.03</td>
<td>17133</td>
</tr>
</tbody>
</table>

-24%
Scenario 3: Dataset Merging

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>@transform</td>
<td>3394 33524</td>
<td>3394 36903</td>
<td>0.42</td>
<td>4.12</td>
<td>8137</td>
<td></td>
<td></td>
</tr>
<tr>
<td>@tbox</td>
<td>3394 36903</td>
<td>&lt;1 4</td>
<td>1.28</td>
<td>13.9</td>
<td>2656</td>
<td></td>
<td></td>
</tr>
<tr>
<td>@smush</td>
<td>3394 36903</td>
<td>3424 38823</td>
<td>0.37</td>
<td>3.98</td>
<td>9265</td>
<td></td>
<td></td>
</tr>
<tr>
<td>@infer</td>
<td>3424 38823</td>
<td>5615 51927</td>
<td>0.32</td>
<td>3.66</td>
<td>10612</td>
<td></td>
<td></td>
</tr>
<tr>
<td>@unique</td>
<td>5615 51927</td>
<td>4085 31297</td>
<td>0.33</td>
<td>3.03</td>
<td>17133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Aggregated</td>
<td>3394 33524</td>
<td>3394 36903</td>
<td>0.41</td>
<td>4.06</td>
<td>8247</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Aggregated</td>
<td>3394 36903</td>
<td>4085 31446</td>
<td>0.14</td>
<td>1.56</td>
<td>23734</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

-24%  -36%
Scenario 4: Dataset Massaging

- TASK: ad-hoc transformations necessary to make data better suited to a particular use
  - data repackaging: preserve data content, but affect the way data is packaged (e.g., changing of RDF syntax)
  - data sanitization: fixing or removing the RDF terms or quads that prevent any further processing of data (e.g., conversion of datatype, URI rewriting, normalisation of literals)
  - data derivation: augmenting a dataset with quads computed from original data (e.g., conversion of a numeric value, counting the occurrences of a certain property for an entity)
- typically implemented in RDFpro using @read, @write and @transform in a single pass without sorting (~0.45 MQ/s)
Evaluation Re-cap

- RDF\textsubscript{pro} implementation of the processing tasks succeeds in managing billions of quads / RDF triples on a commodity machine

- execution times are in the order of hours
  - processing times are negligible if compared to load times in SOA triple stores
    - Virtuoso 7, on same machine, 9h08m for loading 1B triples
  - definitely a winner in one-time processing
Evaluation Re-cap

- RDF\textsubscript{pro} implementation of the processing tasks succeeds in managing billions of quads / RDF triples on a commodity machine

- execution times are in the order of hours
  - processing times are negligible if compared to load times in SOA triple stores
    - Virtuoso 7, on same machine, 9h08m for loading 1B triples
  - definitely a winner in one-time processing

Positively answer our research question!
Conclusions: RDF pro ...

- ... shows that RDF processing tasks on billions of quads can be performed on a single machine using streaming and sorting
- ... a “swiss-army-knife” for exploring and manipulating RDF datasets
- ... is actively used in the NewsReader EU project
- ... is open-source released under the terms of CC0
- ... potentially extendable (future work) to implement restricted versions of OWL 2 inference, SPARQL query answering and SPARQL-based data massaging
RDF_{pro}

Thank you! Questions?

Marco Rospocher
rospocher@fbk.eu :: @marcorospocher
Fondazione Bruno Kessler
Trento, Italy