Implementation and Development Call 20150409

Shortcut to temporal Google doc for meeting notes: http://goo.gl/gvN5ZS

Announcements

- Apps & Tools Working Group: next call April 21 at 1pm ET
  - Link to presentation sign up sheet
  - Link to presentation template
- Ontology Working Group: next call May 5 at 1pm ET
  - When there's no Apps & Tools or Ontology call, there's usually an "office hours" call for informal, unscripted questions about the ontology, but not on next Tuesday 4/14
- 2015 VIVO Conference Call for Papers now open!
  - All submissions through EasyChair due by Friday, April 24th 5:00 PM PST
  - More info and links online: http://vivoweb.org/blog/2015/04/2015-call-papers-now-open
- ORCID/VIVO webinar "Enhancing Early Career Researcher Profiles: VIVO & ORCID Integration" on Thursday, April 16th 11:00 AM EST
  - Additional info and registration link here: http://www.duraspace.org/node/2498
- Site updates on next week's call
- Next theme call on 4/23: Kristi Holmes will be joining to talk about VIVO and other communities
- Tech Lead job description has been posted at http://duraspace.org/jobs

Theme: VIVO on Virtuoso Open-Source Edition

- As reported on last week's call, the Weill Cornell VIVO is now running on Virtuoso Open-Source Edition.
  - "Virtuoso is a scalable cross-platform server that combines Relational, Graph, and Document Data Management with Web Application Server and Web Services Platform functionality."
  - Virtuoso Open-Source Edition Wiki
  - Virtuoso Open-Source Edition GitHub repo
  - Virtuoso Wikipedia page
- Eliza and Paul from Weill Cornell will join us to discuss how their VIVO 1.7 has been modified to replace Apache Jena with Virtuoso. This will be an open discussion, and we've invited others who have explored replacing Jena with other triple store technologies. Congrats to Eliza!
- Paul: describing the context:
  - Why?
    - their VIVO had slow page load times for larger profiles
    - experienced some production downtime as well
  - did a review of triplestore options and decided on Virtuoso which stood out as performing rather well
  - Eliza tested out Virtuoso performance met expectations, then explored how to connect VIVO to Virtuoso over a couple of months, with help from Jim Blake and others
  - still squashing a few remaining bugs, as for example for using the Harvester
- Eliza -- had to modify some of the Virtuoso integration
  - removed some of the union queries to leverage Virtuoso's efficiency advantage
  - some issues still to work on but it's running pretty well
  - the documentation (Eliza's draft link below) will continue to be updated
  - Paul-- what was the process and what were the main hurdles to overcome?
  - Eliza on the process of moving from Jena to Virtuoso:
    - had some issues starting Tomcat on a non-empty database; would take 1+ day to restart; the previous release (7.1) had some issues with starting and stopping Tomcat
    - issue went away with Virtuoso 7.2 which came out in late February
    - a verbose date issue that always went down to the second -- made some changes on the Freemarker templates
    - still some issues with default graph data -- e.g. some users report when trying to add/edit the research overview manually in VIVO, people were seeing duplication of data
      - still getting the duplication even though the triples are in the default graph (kb2)
  - Paul -- what are the differences between Virtuoso and Jena?
    - Virtuoso takes a lot shorter time to load data -- in the past, when loaded the MySQL database with publications, it would take more than an hour; with Virtuoso it's just a little over one minute
      - easier to remove and re-ingest data because it takes so little time
      - deleting the entire HR graph and re-ingesting takes only about 10 seconds
    - Where does Virtuoso store its triples?
      - It has its own database -- binary on disk
      - It backs up its database -- we have a scheduled daily backup and it's pretty small since it's compressed
        - a diff or the full database? the full database
      - When tried to restore it, to be sure it works, it has to decompress
    - What does the admin interface look like?
    - Virtuoso ships with a SPARQL endpoint -- and it ran much faster than Jena SDB
- Some preliminary stats reported last Thursday:
  - Even the largest profiles (2+ MB) load in 6 seconds
  - Compiling and processing a (???) query takes less than one second
  - Most of the load time is in network lag between the database and the app
  - Still slower for a logged-in user, but Eliza is optimistic that she can get some further gains for logged in users
  - Tried an inference on their staging site and wasn't especially snappy -- looking forward to how things perform with VIVO 1.8
  - Main things that modified in configuring Virtuoso:
    - set the memory to 10GB -- more than the Virtuoso tech support said was needed
    - have it configured to track changes to the database
    - configure the port for the SPARQL endpoint
  - Admin interface

        - includes a SPARQL
        - switched slow queries from using unions to using optionals
    - Still some issues with re-inferencing -- occasionally see an exception with updates -- may be Virtuoso-specific syntax
- Possible questions:
    - How did they acceptance or load test this integration?
        - Paul: we loaded individual profile pages (not sure what tools their colleague uses to measure load time) -- also loaded up to 50 profiles simultaneously to test page performance under some server load
        - Alex -- do you have automated acceptance testing specific to Weill Cornell's VIVO?
        - Paul - no automated tools, but a list (script) of use cases that testers work through such as confirming that if you add an honor or award it shows up in someone's profile
    - Don --- Looks like Virtuoso comes with some DBA-friendly tools. Are there ways to get longer-running measures of performance or identifying long-running queries?
        - You can set up an iSQL window to analyze your queries, in an analysis mode that indicates how your queries are doing
            - compilation, execution, and I/O
        - Does this have fairly robust caching mechanisms for execution paths, etc?
            - Eliza - we know it does caching but we haven't done anything to optimize yet
            - Paul -- hunch is that would want to continue to do server-level caching
    - What kinds of performance gains are they seeing in production?
        - for individual profile pages, the load time is between 3 and 6-7 seconds for non-logged in issues
        - there are some pages like the co-author network that are not as fast
        - logged-in users still see a 12-13 second page load time
            - put in an extra flag during ingest to indicate who the data was harvested by -- if the flag is present, VIVO knows it's data outside the default graph so is not intended for editing, and does not even evaluate beyond that flag for editing based on the type of content
            - but can't a flag for data properties -- so still goes out and searches the whole database to see if any statements are present for that data property
            - Brian -- did you try turning off the extra policy that you have about the graph? Eliza -- just modified the code to not go beyond that policy
        - Don -- is there the same concept with Virtuoso of putting commonly queried data on faster disk?
    - Does Virtuoso increase the server resource (CPU, disk space, memory) requirements?
        - Are using a 4-cpu server with over 10GB of memory -- but consultant said Virtuoso was actually using less than 30% of the resource (considerably less than MySQL would)
        - Has the sense that Virtuoso makes fairly minimal demands
    - Did they have to disable any VIVO 1.7 functionality to accomplish this integration?
        - Not really but did modify some of the code -- documented in the document linked at the bottom
        - had to change some of the update syntax -- e.g., from INSERT DATA to INSERT -- also update syntax
        - modified listviewConfigs to comment out UNION queries -- more work to do there
        - known issues (from Eliza's notes on WC wiki)
            - Virtuoso doesn't allow deletion of blank nodes
    - Why not use the commercial Virtuoso offering? Is commercial support available for their open source version?
        - Eliza mentioned an OpenLink consultant -- any more info about them like availability or cost? Consultant responds to questions on the user list also (for free).
        - Paul: We were investigating various options, focused on performance so assumed they'd need the commercial version
        - Consultant said no benefit to commercial version for 10's of millions of triples
        - you can get commercial support -- bronze, silver, etc.
        - there is a community and they have been helpful -- unless very urgent issue (and some consultants take part on the lists)
        - DBpedia and perhaps UniProt use Virtuoso
    - How might their custom code be contributed to add out-of-the-box Virtuoso support to a future VIVO release?
        - Eliza's custom code has already been contributed. Release 1.8 will ship with Virtuoso compatibility, but not with her mods to the custom list views. - Jim
    - Was the Virtuoso Jena connector required? Are ARQ calls being used in VIVO that require the Jena connector or does Virtuoso support ARQ?
        - There is a Jena connector mentioned in the Virtuoso documentation
        - VIVO is still trying to connect to a database for some reason -- Jim guesses that is because in 1.7 the smoke tests were not yet isolated to only run if using SDB
        - Modified the runtime properties file that specifies the database connections
    - Any testing with 1.8 and the inferencing fixes?
        - Inferencing is still a bit slow but we have yet to try with 1.8.
    - Does this support multiple graphs? If so are you placing your data in the default graph or custom graphs?
        - Yes, it does support multiple graphs, and Weill is using 5 different graphs
    - Does Virtuoso have tools to measure performance and other metrics?
        - answered above -- profiling tools
    - Is it ACID compliant? Any problems with concurrency?
        - asked above -- not sure
    - Have you measured time to rebuild (search?) index - Jena/mysql vs. virtuoso?
        - have not done systematic measurements, but Eliza suspects it's about the same time as before; also for re-inferencing
    - Stability?
        - Up except when have forcibly restarted Tomcat -- every night out of habit from the Jena days
        - Virtuoso did go down once -- the log files are getting very big, so may possibly be related
- Don: Can we explore having Duraspace or the VIVO community have a support contract shared -- most of our problems will be similar -- somehow pool resources this way?
- Eliza's documentation (DRAFT)
    - https://dl.dropboxusercontent.com/u/2014679/Virtuoso-Documentation-Draft.pdf
- Benchmarks:
    - http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V7/#comparison
    - http://www.cse.unsw.edu.au/~iwgdm/2012/Slides/Chris.pdf
    - See more here: https://docs.google.com/document/d/1HvqPXk8hBrUk3Nm0in7hyBPsDYVMO5yHMQLA3t7gIwk/edit
- General questions
    - Brigitte -- what do you mean by the default graph? what VIVO comes with?

    - VIVO uses kb2 by default (just 1) for data (Abox)

Notable List Traffic

See the vivo-dev-all archive and vivo-imp-issues archive for complete email threads

Call-in Information

Calls are held every Thursday at 1 pm eastern time – convert to your time at http://www.thetimezoneconverter.com

- Date: Every Thursday, no end date
- Time: 1:00 pm, Eastern Daylight Time (New York, GMT-04:00)
- Meeting Number: 641 825 891

To join the online meeting

- Go to https://cornell.webex.com/cornell/e.php?AT=WMI&EventID=167096322&RT=MiM2
- If requested, enter your name and email address.
- Click "Join".

1. Call in to the meeting:

   1-855-244-8681 (Call-in toll-free number (US/Canada))

   1-650-479-3207 (Call-in toll number (US/Canada))

2. Enter the access code:

   641 825 891 #

3. Enter your Attendee ID:

   8173 #