A Research Collaboratory for Open Source Software Research

Yongqin Gao, Matt van Antwerp, Scott Christley, Greg Madey
Computer Science & Engineering
University of Notre Dame

ICSE - FLOSS 2007
Minneapolis, MN
May 21, 2007

Supported in part by the National Science Foundation, CISE/IIS-Digital Society & Technology, under Grant No. 0222829
Presentation Outline

• Background
• Collaboratories
  • CyberInfrastructure
  • Research Virtual Organizations (RVOs)
• Research data
• Design
• Utilization statistics
• Summary
Background - FLOSS Research

- Current FLOSS research methods include:
  - Simulation studies
  - Surveys/interviews
  - Software analysis
  - Large-scale data collection
  - Statistical analysis
  - Data mining

- The next steps
  - Collaboratories
  - CyberInfrastructure
  - Research Virtual Organizations (RVOs)
    - Such virtual organizations supporting distributed communities go by numerous names: collaboratories, collaboratories, grid communities, science gateways, science portals, and others (NSF, 2007)
Research Collaboratory

• What is a Collaboratory?
  • Precursor to the idea of the NSF CyberInfrastructure (CI)
  • Collection of shared data, information, analytical toolkits and communication technologies
  • A networked organizational form that includes social processes, collaboration techniques and agreements on norms, principles, value, and rules
Research Collaboratories

- ~200 in a ‘05 taxonomy (http://www.scienceofcollaboratories.org/)
- Bioinformatics - Genomic resources (data & tools)
  - NCBI, FlyBase, Ensembl, VectorBase, WormBase, etc.
- NEES - Network for Earthquake Engineering Simulation
  - NEES is a shared national network of 15 experimental facilities, collaborative tools, a centralized data repository, and earthquake simulation software, all linked by the ultra-high-speed Internet2 connections of NEESgrid.
  - These resources support collaboration and discovery in the form of more advanced research based on experimentation and computational simulations of the ways buildings, bridges, utility systems, coastal regions, and geomaterials perform during seismic events.
Research Collaboratories (cont)

• CLEANER
  – An environmental cyberinfrastructure that provides data archives, collaboration and networking among community members, and information technology for engineering modeling, analysis, and visualization of data
  – Includes a CyberCollaboratory: a collaborative space where communities of researchers, practitioners, and policy-makers, and others come together to share knowledge and information, analyze data, solve problems, and collaborate on publications.
  – The CLEANER Project uses the CyberCollaboratory to support over 100 researchers and educators.
Research Collaboratories (cont)

• FLOSS Research - examples include:
  – FLOSSmole
    • “Screen scraped” data from SourceForge, FreshMeat, RubyForge, FSF, etc.
  – CVSAnalY - GSyC/LibreSoft
    • CVS/Subversion statistical analysis tool
  – The SourceForge Research Data Data Archive
    • Archive of SourceForge.net back-end database dumps
    • Wiki-based collaboratory
    • This presentation!
Research Data Description

• SourceForge.net
  – The largest OSS development community
  – 148,000+ registered projects
  – 1,586,000+ registered users
  – Project data
    – Downloads, bug reports, forum activity, developers, project characteristics, etc.
  – Developer data
    – Activity
    – Project membership
Research Data Description

• Our Data Set
  – 488G total and growing at 12G/month.
  – Every dump has 80-120 tables.
  – Tables have up to 30 million records.

• Hosting Environment
  – Dual Xeon 3.06GHz, 4G RAM, 2T RAID storage
  – Linux 2.4.21-40.ELsmp with PostgreSQL 8.1
Design

• Presentation Tier
  • Browser interface
  • Wiki

• Logic Tier
  • Authentication
  • Schema browser
  • Queries & download

• Data Tier
  • PostgreSQL
  • Monthly schema
Data Tier

• PostgreSQL
• Database - timeline
• Monthly schema: one for each dump
• Mirrors the SourceForge.net backend

Every schema is a database dump from the SourceForge.net
Data Tier

- Connection pool
- Persistent connections for improved performance
Presentation Tier

- Various access methods
- Documentation and references
- Community support - FAQ, schema browser, table definitions
- Wiki interface
Schema Browser

Sf0407

Schema sf0407 has 83 tables.

- pg_stat_database_historical
- trove_egg
- user_perms
- groups_registration
- pg_autovac_skip
- ref_timezones
- stats_outage_log
- stats_rank_oldformula_byday
- stats_fileid_alltime_egg
- stats_group_rank_alltime
- stats_group_rank_byday
- stats_groupid_alltime_egg
- stats_toplist_week
- trove_cat_activity
- trove_ref_translation_to_iso8859
- trove_egg_minix
- artifact
- artifact_canned_responses
- artifact_category
- artifact_counts_egg
- artifact_file

Schemas

- January 2003 - sf0103
- November 2004 - sf1104
- December 2004 - sf1204
- February 2005 - sf0205
- March 2005 - sf0305
- April 2005 - sf0405
- May 2005 - sf0505
- June 2005 - sf0605
- July 2005 - sf0705
- August 2005 - sf0805
- September 2005 - sf0905
- October 2005 - sf1005
- November 2005 - sf1105
- December 2005 - sf1205
- January 2006 - sf0106
- February 2006 - sf0206
- March 2006 - sf0306
Stats project all

Appears in the following schemas:

- sfo103
- sfo1104
- sfo1204
- sfo205
- sfo305
- sfo405
- sfo505
- sfo605
- sfo705
- sfo805
- sfo905
- sfo1005
- sfo1105
- sfo1205
- sfo10106
- sfo10206
- sfo10306
- sfo10406
- sfo10506
- sfo10606
- sfo10706
- sfo10806
- sfo10906
- sf1006
- sf1106
- sf1206
- sf10107
- sf10207
- sf10307

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Modifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>group_id</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>developers</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>group_ranking</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>group_metric</td>
<td>double precision</td>
<td></td>
</tr>
<tr>
<td>logo_showings</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>downloads</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>site_views</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>subdomain_views</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>page_views</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>msg_posted</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>msg_uniq_auth</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>bugs_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>bugs_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>support_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>support_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>patches_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>patches_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>artifacts_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>artifacts_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>tasks_opened</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>tasks_closed</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>help_requests</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>cvs_checkouts</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>cvs_commits</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>cvs_adds</td>
<td>integer</td>
<td></td>
</tr>
<tr>
<td>svn_checkouts</td>
<td>integer</td>
<td></td>
</tr>
</tbody>
</table>
SourceForge.net Research Archive Query Form

Examples

SELECT *
FROM sf0305.users
WHERE user_id < 100

SELECT user_name
FROM sf1104.users a,
    sf1104.artifact b
WHERE a.user_id =
b.submitted_by AND
b.artifact_id = 304727

SELECT:

FROM:

WHERE:

Separator
- ;
- ,
- #
- XML

Add SQL query to result file?
- yes
- no

Submit Query  Clear

News

- The database version has been upgraded. If you notice any errors, please let us know (oss at nd dot edu)
- SQL query option added (as an attribute of the root element in XML output, as the first line in text file output)
- April schema (sf0407) now loaded.
Logic Tier

• Interactive web query system
  – Authorized user can submit query to the back end repository through the web query
  – Results are provided by files with various formats: text with various delimiters, XML

– Dynamic web schema browser
  – Authorized user can access the dynamic schema of the repository through the schema browser
Utilization - Sample

• Monthly activity (June 2006)
  – Total queries submitted: \(16,947\)
  – Total data files retrieved: \(13,343\)
  – Total bytes of query data downloaded: \(26,684,556,278\)

• Monthly activity (Feb 2007)
  – Total queries submitted: \(38,659\)
  – Total data files retrieved: \(24,422\)
  – Total bytes of query data downloaded: \(13,048,335,165\)
Utilization Statistics - 2006

Queries and Downloads by Month

Month of 2006

Number of Queries

Gigabytes Downloaded

January | February | March | April | May | June | July | August | September | October | November

0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0

3,000 | 4,000 | 5,000 | 6,000 | 7,000 | 8,000 | 9,000 | 10,000 | 11,000 | 12,000 | 13,000

4,000 | 5,000 | 6,000 | 7,000 | 8,000 | 9,000 | 10,000 | 11,000 | 12,000 | 13,000 | 14,000

0 | 5 | 10 | 15 | 20 | 25 | 30

queries | gigabytes downloaded
Summary

• SourceForge.net archive
  – http://zerlot.cse.nd.edu/
  – Access open to all academic/scholarly researchers - sublicense

• Plans
  – Programmable access method should be provided for complicated access
    – Web services in testing phase
  – Analysis/data mining tools, preselected data sets, etc.
  – FLOSS CyberInfrastructure - network of collaboratories?
  – FLOSS Research Virtual Organization (FLOSS-RVO)
Thank You!