Ultimate Debian Database

Lucas Nussbaum

debian

Debconf 9
Debian: the data hell

A lot of different sources of data in Debian

With different data formats:
text files, BerkeleyDB, SQL databases, ...

Need to combine them all:
Mainly for Quality Assurance, e.g:

- Packages of priority $\geq$ standard with RC bugs?
- Maintainers with lots of outdated/buggy packages?
Idea:

- **Import all the data in a single (PostgreSQL) SQL DB**
- Easier to query (relatively well-known interface)
- *The* proper way of joining data together
- No need to write problem-specific scripts
  - e.g. the *Popular orphaned packages* ?
History

Started as a **Google Summer of Code** project in 2008

Student: Christian von Essen (Neronus)

Mentors:
- Lucas Nussbaum (lucas)
- Marc Brockschmidt (HE)
- Stefano Zacchiroli (zack)

Results:
- Very good work from Christian
- Usable code at the end of the summer
  - mostly Python, some Perl
Design choices

Not problem-specific, no typical queries
(not projectb or the new wanna-build DB !)

Schema :

- Typical user == human
- Make it easy to write/run queries
- Performance ? important, but not a critical goal
- No surrogate keys
Surrogate key

- Unique identifier (usually integer)
- Used as primary key
- Not derived from any application data

packages (\texttt{package\_id}, \texttt{package\_name}, ...)

MySQL : AUTO\_INCREMENT
PostgreSQL : serial

Has both advantages and disadvantages

Details : http://en.wikipedia.org/wiki/Surrogate\_key
Design choices (2)

Data:
- Correctness is critical
- Partial updates? Often difficult/risky
- Solution: complete data reloads
  - Using transactions to avoid temporary unavailability
Debian is inconsistent
- What does "package" mean?

Inconsistency can be interesting for QA
⇒ Keep inconsistency in UDD
⇒ No foreign keys between data sources
Current status

- Hosted on udd.debian.org (dedicated machine)
- Uses PostgreSQL 8.4
- You can connect from {merkel, alioth, master}.d.o
  - e.g.: /usr/lib/postgresql/8.4/bin/psql
    service=udd
- Even non-DDs can connect!

More info:
http://wiki.debian.org/UDD
What we currently import

Sources and Packages
Bugs (including archived bugs)
Carnivore
Debtags
Popularity Contest
DEHS (upstream status)
Debian LDAP (restricted)
Lintian

Migrations to testing
History of uploads
NEW queue
DDTP (translation status)
Orphaned packages
Screenshots
Ubuntu Sources/Packages
Ubuntu bugs

Imported using:

- Working and monitored scripts
- Ran regularly (cron jobs or ssh triggers)
So, what can we find out about Debian using UDD?
Source-only uploads, anyone?

- Upload without any architecture-specific package
  - only source and arch: all packages
- Already possible!
- Who does it?

Uploads:
- Of arch: any source packages
- With only arch: all packages
Source only uploads, anyone? (2)

SELECT package, version, signed_by
FROM upload_history uh
WHERE package in
  (SELECT source FROM sources
   WHERE distribution = 'debian' AND release = 'sid'
     AND architecture = 'any')
AND NOT EXISTS
  (select * from upload_history_architecture uha
   WHERE uh.id = uha.id
   AND uha.architecture NOT IN ('all','source'))
ORDER BY date DESC;
### Source only uploads, anyone? (3)

<table>
<thead>
<tr>
<th>package</th>
<th>signed_by_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>zsh-beta</td>
<td>Clint Adams</td>
</tr>
<tr>
<td>eglibc</td>
<td>Aurelien Jarno</td>
</tr>
<tr>
<td>perl</td>
<td>Niko Tyni</td>
</tr>
<tr>
<td>git-core</td>
<td>Gerrit Pape</td>
</tr>
<tr>
<td>ptlib</td>
<td>Mark Purcell</td>
</tr>
<tr>
<td>ppl</td>
<td>Arthur Loiret</td>
</tr>
<tr>
<td>git-core</td>
<td>Gerrit Pape</td>
</tr>
<tr>
<td>alarm-clock</td>
<td>Piotr Ozarowski</td>
</tr>
<tr>
<td>xtables-addons</td>
<td>Pierre Chifflier</td>
</tr>
<tr>
<td>coccinelle</td>
<td>Eugeniy Meshcheryakov</td>
</tr>
<tr>
<td>dulwich</td>
<td>Jelmer Vernooij</td>
</tr>
<tr>
<td>libparams-classify-perl</td>
<td>Rene Mayorga</td>
</tr>
<tr>
<td>kaya</td>
<td>Stuart Teasdale</td>
</tr>
<tr>
<td>eigen2</td>
<td>Ana Beatriz Guerrero Lopez</td>
</tr>
</tbody>
</table>
select package, count(distinct tag) as cnt 
from lintian 
where tag_type in ('error','warning') 
group by package 
order by cnt desc limit 15;
Number of different lintian errors or warnings

```sql
select package, count(distinct tag) as cnt 
from lintian
where tag_type in('error','warning')
group by package
order by cnt desc limit 15;
```

<table>
<thead>
<tr>
<th>package</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>openswan</td>
<td>23</td>
</tr>
<tr>
<td>heartbeat</td>
<td>21</td>
</tr>
<tr>
<td>tcpquota</td>
<td>18</td>
</tr>
<tr>
<td>replicator</td>
<td>17</td>
</tr>
<tr>
<td>nws</td>
<td>17</td>
</tr>
<tr>
<td>hercules</td>
<td>17</td>
</tr>
<tr>
<td>harden-doc</td>
<td>17</td>
</tr>
<tr>
<td>euro-support</td>
<td>17</td>
</tr>
</tbody>
</table>
Let’s look at more positive things!
Who uploaded sid’s packages?

```sql
select changed_by, count(*) from sources s, upload_history uh
where s.source = uh.source and s.version = uh.version
and s.distribution='debian' and s.release = 'sid'
group by changed_by order by count desc limit 8;
```
Who uploaded sid’s packages?

```sql
select changed_by, count(*) from sources s, upload_history uh
where s.source = uh.source and s.version = uh.version
and s.distribution='debian' and s.release = 'sid'
group by changed_by order by count desc limit 8;
```

<table>
<thead>
<tr>
<th>changed_by</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel Baumann <a href="mailto:daniel@debian.org">daniel@debian.org</a></td>
<td>208</td>
</tr>
<tr>
<td>gregor herrmann <a href="mailto:gregoa@debian.org">gregoa@debian.org</a></td>
<td>196</td>
</tr>
<tr>
<td>Ryan Niebur <a href="mailto:ryanryan52@gmail.com">ryanryan52@gmail.com</a></td>
<td>196</td>
</tr>
<tr>
<td>Barry deFreese <a href="mailto:bdefreese@debian.org">bdefreese@debian.org</a></td>
<td>163</td>
</tr>
<tr>
<td>Torsten Werner <a href="mailto:twerner@debian.org">twerner@debian.org</a></td>
<td>139</td>
</tr>
<tr>
<td>Dirk Eddelbuettel <a href="mailto:edd@debian.org">edd@debian.org</a></td>
<td>127</td>
</tr>
<tr>
<td>Christian Perrier <a href="mailto:bubulle@debian.org">bubulle@debian.org</a></td>
<td>126</td>
</tr>
<tr>
<td>gregor herrmann <a href="mailto:gregor+debian@comodo.priv.at">gregor+debian@comodo.priv.at</a></td>
<td>123</td>
</tr>
</tbody>
</table>
Who uploaded sid’s packages?

Using carnivore:

```sql
select cn.name, count(*)
from sources s, upload_history uh,
carnivore_emails ce, carnivore_names cn
where s.source = uh.source and s.version = uh.version
and s.distribution='debian' and s.release = 'sid'
and changed_by_email = ce.email and ce.id = cn.id
group by cn.name order by count desc limit 20;
```
Who uploaded sid’s packages?

Using carnivore:

```
select cn.name, count(*)
from sources s, upload_history uh,
carnivore_emails ce, carnivore_names cn
where s.source = uh.source and s.version = uh.version
and s.distribution='debian' and s.release = 'sid'
and changed_by_email = ce.email and ce.id = cn.id
group by cn.name order by count desc limit 20;
```

<table>
<thead>
<tr>
<th>name</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gregor Herrmann</td>
<td>319</td>
</tr>
<tr>
<td>Barry deFreese</td>
<td>211</td>
</tr>
<tr>
<td>Daniel Baumann</td>
<td>208</td>
</tr>
<tr>
<td>Ryan Niebur</td>
<td>196</td>
</tr>
<tr>
<td>Torsten Werner</td>
<td>139</td>
</tr>
<tr>
<td>Dirk Eddelbuettel</td>
<td>127</td>
</tr>
<tr>
<td>Christian Perrier</td>
<td>126</td>
</tr>
</tbody>
</table>
Who reported lenny’s RC bugs?

RC bugs reported since the release of etch (08/04/07)

select submitter_name, count(*) from all_bugs
where status = 'done' and arrival >= '2007-04-08'
and severity >= 'serious'
and submitter_name != ''
group by submitter_name
order by count desc limit 10;

<table>
<thead>
<tr>
<th>Name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucas Nussbaum</td>
<td>2287</td>
</tr>
<tr>
<td>Kurt Roeckx</td>
<td>528</td>
</tr>
<tr>
<td>Bastian Blank</td>
<td>481</td>
</tr>
<tr>
<td>Michael Ablassmeier</td>
<td>462</td>
</tr>
<tr>
<td>Daniel Schepler</td>
<td>445</td>
</tr>
<tr>
<td>Matthias Klose</td>
<td>339</td>
</tr>
<tr>
<td>Frank Lichtenheld</td>
<td>271</td>
</tr>
<tr>
<td>Nico Golde</td>
<td>209</td>
</tr>
</tbody>
</table>
Who reported lenny’s RC bugs?

RC bugs reported since the release of etch (08/04/07)

```sql
select submitter_name, count(*) from all_bugs
where status = 'done' and arrival >= '2007-04-08'
and severity >= 'serious'
and submitter_name != ''
group by submitter_name
order by count desc limit 10;
```

<table>
<thead>
<tr>
<th>name</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lucas Nussbaum</td>
<td>2287</td>
</tr>
<tr>
<td>Kurt Roeckx</td>
<td>528</td>
</tr>
<tr>
<td>Bastian Blank</td>
<td>481</td>
</tr>
<tr>
<td>Michael Ablassmeier</td>
<td>462</td>
</tr>
<tr>
<td>Daniel Schepler</td>
<td>445</td>
</tr>
<tr>
<td>Matthias Klose</td>
<td>339</td>
</tr>
<tr>
<td>Frank Lichtenheld</td>
<td>271</td>
</tr>
<tr>
<td>Nico Golde</td>
<td>209</td>
</tr>
</tbody>
</table>
What was used in those examples

**Sources and Packages**
- Bugs (including archived bugs)
- Carnivore
- Debtags
- Popularity Contest
- DEHS (upstream status)
- Debian LDAP (restricted)
- Lintian

**Migrations to testing**
- History of uploads
- NEW queue
- DDTP (translation status)
- Orphaned packages
- Screenshots
- Ubuntu Sources/Packages
- Ubuntu bugs
UDD users

- "Ubuntu" box on the PTS
  - You don’t want to hear about the Launchpad bugs importer
- Debian Data Export (Enrico Zini)
- HELIOS project (Olivier Berger) : database of facts about Free Software projects
- Debian Pure Blend (Andreas Tille, next talk)
- Bapase
Your turn

- SSH merkel.debian.org (sitting next to samosa)
  or SSH alioth.debian.org

- Connect to UDD:
  ```
  /usr/lib/postgresql/8.4/bin/psql service=udd
  ```

- Look around:
  ```
  \dt
  \dt tablename
  ```

- Run your first query:
  ```
  select * from sources
  where maintainer_email = '<your email>';
  ```
Let’s look at orphaned packages

(they don’t get enough attention anyway)

```sql
select * from orphaned_packages;
```
Let’s look at orphaned packages

(they don’t get enough attention anyway)

```sql
select * from orphaned_packages;
select count(*) from orphaned_packages
where type != 'RFA';
```

Lucas Nussbaum   Ultimate Debian Database
Let’s look at orphaned packages

(they don’t get enough attention anyway)

```sql
select * from orphaned_packages;
select count(*) from orphaned_packages
where type != 'RFA';
select * from popcon_src;
```
Let’s look at orphaned packages

(they don’t get enough attention anyway)

```sql
select * from orphaned_packages;
select count(*) from orphaned_packages where type != 'RFA';
select * from popcon_src;
select insts, o.source, description
from orphaned_packages o, popcon_src p
where o.source = p.source
order by insts desc;
```
Future work / Help needed!

Help needed:

- Play with UDD, build tools on top of the DB
  - So we know what’s missing/should be improved
- Implement missing importers
  - wanna-build, britney, MIA
- Improve examples and documentation

Contact: #debian-qa or debian-qa@l.d.o

http://wiki.debian.org/UDD