Ultimate Debian Database

Lucas Nussbaum

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- *Maintainers with lots of outdated/buggy packages ?*

Ultimate Debian Database

Idea :

- **Import all the data in a single (Postgre)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 (**package_id**, 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

Design choices (3)

Debian is inconsistent

- What does "package" mean ?

Inconsistency can be interesting for QA

$\Rightarrow$ Keep inconsistency in UDD
$\Rightarrow$ 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)

package	signed_by_name
zsh-beta	Clint Adams
eglibc	Aurelien Jarno
perl	Niko Tyni
git-core	Gerrit Pape
ptlib	Mark Purcell
ppl	Arthur Loiret
git-core	Gerrit Pape
alarm-clock	Piotr Ozarowski
xtables-addons	Pierre Chifflier
coccinelle	Eugeniy Meshcheryakov
dulwich	Jelmer Vernooij
libparams-classify-perl	Rene Mayorga
kaya	Stuart Teasdale
eigen2	Ana Beatriz Guerrero Lopez

Number of different lintian errors or warnings

```
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

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

package	count
openswan	23
heartbeat	21
tcpquota	18
replicator	17
nws	17
hercules	17
harden-doc	17
euro-support	17

Let’s look at more positive things !

Who uploaded sid's packages ?

```
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 ?

```
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;
```

changed_by	count
Daniel Baumann <daniel@debian.org>	208
gregor herrmann <gregoa@debian.org>	196
Ryan Niebur <ryanryan52@gmail.com>	196
Barry deFreese <bdefreese@debian.org>	163
Torsten Werner <twerner@debian.org>	139
Dirk Eddelbuettel <edd@debian.org>	127
Christian Perrier <bubulle@debian.org>	126
gregor herrmann <gregor+debian@comodo.priv.at>	123

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;
```

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;
```

name	count
Gregor Herrmann	319
Barry deFreese	211
Daniel Baumann	208
Ryan Niebur	196
Torsten Werner	139
Dirk Eddelbuettel	127
Christian Perrier	126

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;
```

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;
```

name	count
Lucas Nussbaum	2287
Kurt Roeckx	528
Bastian Blank	481
Michael Ablassmeier	462
Daniel Schepler	445
Matthias Klose	339
Frank Lichtenheld	271
Nico Golde	209

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)

```
select * from orphaned_packages;
```

Let's look at orphaned packages

(they don't get enough attention anyway)

```
select * from orphaned_packages;
```

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

Let's look at orphaned packages

(they don't get enough attention anyway)

```
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)

```
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`