Postgres, Collations, And File System Level Copies

Matthew Kelly

TripAdvisor

The Context

- Every file system based copy (physical streaming, WAL archiving, pg_basebackup, rsync, etc.) copies indexes around in sorted order.

- Postgres relies on the system to implement collation support. It is up to the host system to decide what en.utf-8 means.

The Problem

- Collations are not stable between versions of glibc. Ex. CentOS (RHEL) 5.10 vs. 6.4

```
[mkelly@bro01a ~]$ /usr/local/pgsql/bin/psql -U tripmaster -h dbr05a
psql (9.1.9)
Type "help" for help.

tripmaster=# SELECT 'RÜSSEL' > 'RUßEL';
 ?column?
_____
 t
(1 row)

tripmaster=# \q
[mkelly@bro01a ~]$ /usr/local/pgsql/bin/psql -U tripmaster -h dbr06a
psql (9.1.9)
Type "help" for help.

tripmaster=# SELECT 'RÜSSEL' > 'RUßEL';
 ?column?
_____
 f
(1 row)
```

Manifestations at TripAdvisor

- Single query on an inner join
  - Merge join: 0 Rows
  - Hash Join: ~50,000,000 rows
- Found on 8 internal production, and 2 live site servers
- Up to 3771 rows out of place in a single index
- Up 12806 rows out of place across 26 indexes on one machine
- On average roughly 15% of indexes containing text keys on tables larger 100MB were found to exhibit this issue
- In at least one case, rebuilding a unique index on a master revealed that the database had allowed 100+ primary key violations.

Try it Your Self

[https://github.com/mkellycs/
postgres_index_integrity_check](https://github.com/mkellycs/postgres_index_integrity_check)

There is a current discussion on hackers about the future options for those who are interested in more detail.