

Indexování XML v databázi PostgreSQL

Tomáš Pospíšil

www.tomaspospisil.com

9. února 2012

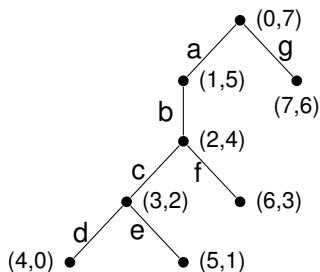


- SELECT/UPDATE/DELETE nad strukturovaným datovým typem
- XPath/XQuery v DB
- Datově orientovaná XML vs dokumentově orientovaná XML
- DB jako úložiště XML dokumentů
- Sémantika XML dat (validace oproti schématům)
- Nativní XML DB x ORDMS XML podpora

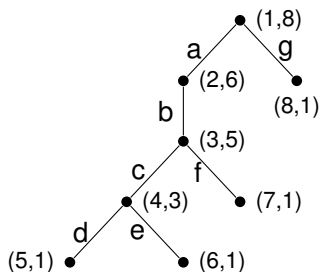


- XML datový typ
 - Uložení XML dokumentů (CLOB)
 - Indexace pouze pomocí XPATH dotazu
 - SQL/XML funkce (xmlcomment, xmlconcat, ...)
 - XML2 extension
 - LibXML 2 knihovna
- Chybějící podpora
 - Indexace XML
 - XML validace pomocí schémat DTD, XSD a RNG
 - XPath 2.0, XQuery
 - Datové typy a operace





- Stromová struktura (*pre_order*, *post_order*)
- Každá aktualizace vyžaduje přepočít



- Každý uzel je určen $(order, size)$
- $order(x) < order(y)$ pro $\forall$ synovské uzly y uzlu x
- $order(x) + size(x) \geq order(y) + size(y)$
- $size$ lze volit libovolně, při splnění podmínek

Element table
Document_ID
Order
Size
Tab_Name
Depth
Child_ID
Next_ID
Attr_ID

Attributes table
Document_ID
Order
Size
Tab_Name
Depth
Parent_ID
Next_ID
Value

Text table
Document_ID
Order
Size
Depth
Parent_ID
Next_ID
Value

Document table
Document_ID
Name

- `xml_element_nodes` table

- `xml_element_nodes_pkey` PRIMARY KEY, btree (did, pre_order, size)
- `elem_tab_all_index` btree (name, did, pre_order, size)
- `elem_tab_range_index` gist (int4range(pre_order, pre_order + size))

- `xml_attribute_nodes` table

- `xml_attribute_nodes_pkey` PRIMARY KEY, btree (did, pre_order)
- `attr_tab_all_index` btree (name, did, pre_order)

- `xml_text_nodes` table

- `xml_text_nodes_pkey` PRIMARY KEY, btree (pre_order, did)
- `text_tab_index` btree (parent_id, did)

- xml2 extension
- Jazyk C, PL/pgSQL
- Použití in4range (vhodná operace <@ "is contained by")
- Nové datové typy DTD, XSD a RNG v PostgreSQL
- API knihovny LibXML 2



- Funkční modul v *contrib/xml2* rozšíření (README.txt)
- Dokumentace kódu v rámci oficiální dokumentace (SGML)
- Instalace
 - `./configure --with-libxml --with-libxslt`
 - `make contrib/xml2; make install`
 - `psql; CREATE EXTENSION xml2`
- <http://www.tomaspospisil.com>
- [git://github.com/killteck/indexing-xml.git](https://github.com/killteck/indexing-xml.git)



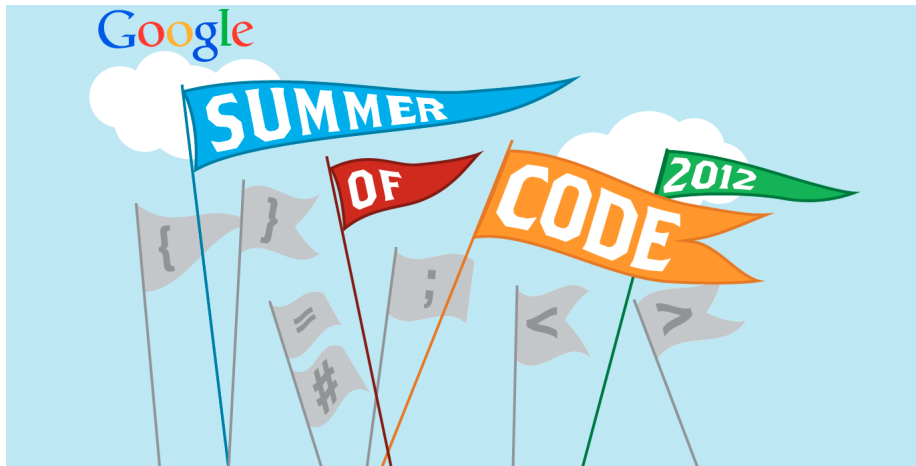
XML	DTD (ms)	XML	XSD (ms)	XML	RNG (ms)
1	32,868	1	23,110	1	1,894
2	20,134	2	25,801	2	1,451
3	1,314	3	32,234	3	31,898
4	40,030	4	26,045	4	32,266
5	1,410	5	1,868	5	30,109

Tabulka: Časy validace pomocí DTD, XSD a RNG nad XML dokumenty cca 1 MB

- Rychlost validace v řádu ms
- Paměťová náročnost lineární (velikost XML dokumentu + validačních schémat)
- Core2Duo T9800 @ 2.93GHz, 4GB RAM, Hitachi 7200rpm@ext3



- Projekt implementovaný v rámci GSoC 2011
- Využití indexu pro XPath/SQL dotazování
- Návrh na nativní XML podporu, využití indexu
- Mentor Gregory Stark



- 8. ročník
- 5000 \$
- Výborná pracovní zkušenost (10 nabídek za 1 hod.)
- Spousta TODO v PostgreSQL

