doi:10.3359/oz1314042

poročilo

KONFERENCA *LUCENE/SOLR REVOLUTION 2013*

UVOD

Lucene/Solr Revolution je največja odprtokodna konferenca, posvečena Apachejevemu projektu Lucene/Solr. Konferenco je organiziralo podjetje LucidWorks, ki zaposluje dobro četrtino vseh razvijalcev projekta Lucene/Solr in tudi zagotavlja podporo za omenjena projekta.

Konferenca je potekala od 4. do 7. novembra v Dublinu, glavnem mestu Irske, ki je tudi ekonomsko, administrativno in kulturno središče otoka. Leži na vzhodni obali ob izlivu reke Liffey v Irsko morje. Mesto, ki si ga ustanovili Vikingi, ima danes okoli pol milijona prebivalcev. Konferenca je potekala v mestni četrti Ballsbridge, poimenovani po mostu preko reke Dodder, in sicer na nacionalnem stadionu, kjer igrata irski izbrani vrsti nogomet in ragbi.

Konferenca je potekala v dveh delih: prvi je trajal dva dneva in je bil namenjen praktičnemu tečaju, v drugem delu pa so potekala predavanja.

KAJ JE SPLOH LUCENE/SOLR

Apache Lucene je odprtokodni zmogljiv iskalnik, napisan v celoti v jeziku Java, z naslednjimi lastnostmi:

- dostop preko enostavnega API-ja,
- hitro indeksiranje,
- nizke strojne zahteve,
- zmogljivi, natančni ter učinkoviti iskalni algoritmi,
- rangirano iskanje,
- različne vrste iskanja (frazno iskanje, iskanje s krajšanjem, iskanje po območju, iskanje z bližino …),
- razvrščanje rezultatov,
- fasetiranje, grupiranje.

Apache Solr je iskalna platforma nad Lucenom. Glavne lastnosti so: polno iskanje po besedilih, označevanje zadetkov, fasetno iskanje, indeksiranje skoraj v realnem času, integracija podatkovne baze, rokovanje z dokumenti (word, PDF) in geoprostorsko iskanje. Solr je zelo zanesljiv, nadgradljiv ter toleranten na napake. Omogoča porazdeljeno indeksiranje, podvajanje strežnikov ter izenačevanje obremenitve pri iskanjih. Do Solr-ja je mogoče dostopati preko http/xml in programskih vmesnikov JSON.

PRAKTIČNO USPOSABLJANJE

Praktično usposabljanje je potekalo v treh različnih tečajih, ki so potekali istočasno: Solr Unleashed, Solr Under the Hood in Big Data & Solr. Udeležil sem se tečaja Solr Unleashed.

Big Data & Solr je bil namenjen razvijalcem, ki želijo:

- vedeti več o ključnih odprtokodnih orodjih, kot so Hadoop, Cascading in Mahout,
- procesirati masivne podatke in generirati velike iskalne indekse,
- uporabljati Solr kot nadgradljivo bazo NoSQL.

Solr Under the Hood je bil namenjen vsem, ki Solr že poznajo, a želijo vedeti, kako le-ta deluje in kako maksimalno izkoristiti njegovo zmogljivost.

Solr Unleashed

Na začetnem tečaju so predstavili vse pomembne funkcije Solr in odgovorili na vsa vprašanja, ki se pojavijo pri razvoju iskalnika; recimo, kaj je treba spremeniti, da bi lahko bolje izkoristili zmogljivosti Solr.

Tečaj sestavlja 8 logičnih sklopov in 25 praktičnih nalog.

1. Osnove

Prikazali so nametitev Solr, dodajanje vsebine v Solr in osnovno iskanje. Iščemo lahko preko URL API-ja (primer: http://localhost:8983/solr/select/?q=Cankar) ali pa preko vmesnika za brskanje, ki je običajno namenjen testiranju in se ne uporablja v produkciji.

2. Iskanje

Prikazali so sortiranje rezultatov, različne razčlenjevalnike poizvedb (angl. *query parsers*), fiksiranje parametrov iskanja, fasetno iskanje in grupiranje rezultatov iskanja.

3. Indeksiranje
Začeli smo graditi iskalno aplikacijo za trgovino s knjigami, poudarek pa je bil na različnih vrstah podatkov, dinamičnih in statičnih iskalnih poljih ter dodajanju in brisanju podatkov.

4. Shema Solr
Vse o posameznih iskalnih poljih, vrstah podatkov in obdelavi teh podatkov, preden se shranijo v indeks, se nastavi v shemi Solr. Predstavili so spreminjanje te sheme in različne analizatorje, filtre ter žetone besedila.

5. Relevantnost zadetkov
Da kot uporabnik dobimo pričakovane zadetke, lahko uporabimo obtežitev iskalnih polj, frazno iskanje, funkcijsko iskanje, nadomestne znake, mehko iskanje ter fonetično iskanje (angl. *sounds-like*). Na žalost Solr nima posebej dobre podpore za fonetično iskanje v slovanskih jezikih.

6. Napredne funkcije
Predstavljene so bile funkcije, kot so: več podobnih zadetkov (angl. *more-like this*), geoprostorsko iskanje, preverjanje črkovanja, predlogi (angl. *auto complete*), označevanje zadetkov, večjezično iskanje, navidezna polja in navidezno združevanje.

7. Jedra Solr
Predstavljeno je bilo upravljanje strežnika Solr z več jedri (dodajanje, brisanje jeder, zakaj sploh uporabljati jedra).

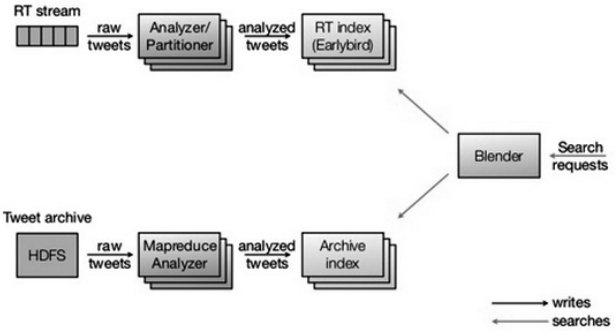
8. SolrCloud
SolrCloud so nove porazdelitvene kapacitete Solr, ki omogočijo avtomatsko distribucijo pri indeksiranju, porazdeljeno iskanje v razdeljenem okolju, avtomatsko dodajanje podatkovnih kopij. Spoznali smo, kako postaviti takšno okolje s pomočjo ZooKeeperja, ki skrbi za konfiguracijo postavitve in sinhronizacijo med posameznimi strežniki Solr.

KONFERENCA

V nadaljevanju bomo na kratko predstavili Twitter ter nekaj predavanj, ki so najbolj relevantna za razvoj novega iskalnika na IZUM-u. Videoposnetki predavanj in predstavitve so javno dostopni na http://www.lucenerevolution.org/2013/Lucene-Solr-Revolution-2013-Dublin-Presentations.

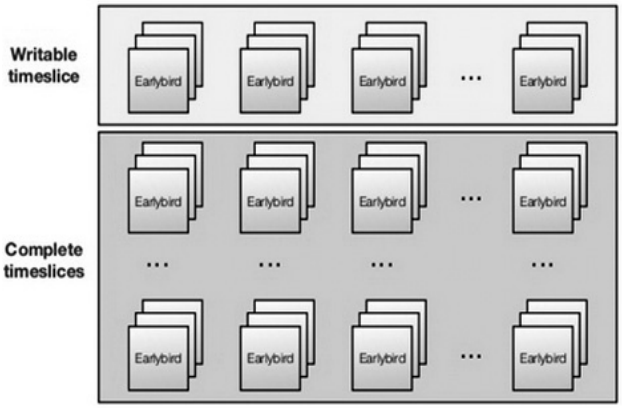
Twitter

Gre za spletno družbeno omrežje in mikroblogno storitev, ki svojim uporabnikom omogoča, da med seboj izmenjujejo kratka sporočila, dolga do 140 znakov. Twitterjev iskalnik mora izvesti dve milijardi iskanj na dan, saj ima Twitter več kot 230 milijonov aktivnih mesečnih uporabnikov, ki pošljejo 500 milijonov tvitov na dan.



Slika 1: Arhitektura Twitterja

Za zagotovitev iskanja v realnem času uporablja Twitter prilagojen iskalnik Lucene. Posebnost je, da lahko iščemo po medpomnilniku IndexWritter, to pa zato, ker so tviti fiksne dolžine in ker je iskanje lahko razvrščeno samo od najnovejših proti starejšim. Ko se IndexWritter napolni, se odpre nov IndexWritter, napolnjen pa se zaklene in ni več na voljo za pisanje. Iskanje poteka od konca IndexWritterja proti začetku in ko je najdenih dovolj rezultatov, se zaključi.



Slika 2: Arhitektura indeksiranja

Indeks arhiv je standardni indeks Lucene, razvrščen časovno. Razdeljen je na dva dela, na del v pomnilniku, ki vsebuje najboljše tvite, ter na del na disku, ki vsebuje vse tvite. Iskanje po disku se izvede samo, če iskanje po pomnilniku ne da dovolj rezultatov.

The Typed Index

Zastavljeno je bilo vprašanje, kako narediti dober iskalnik v večjezičnem okolju, pri čemer namesto različnih iskalnih polj ali celo celih indeksov raje dodamo tip podatka kot prepono k indeksiranim nizom. Razložili so, kako se izvede iskanje po takem indeksu in kako se uporablja funkcija "SpanQuery" za frazno in fonetično iskanje.

SpellChecking in Trovit

Predstavljeno je bilo preverjanje črkovanja v primeru iskalnika Trovit, ki je večjezična iskalna platforma za oglase. Pri tem se uporablja mešan sistem splošnih slovarjev in slovarjev, narejenih iz dejanskega indeksa. S tem dobijo slovnično pravilno črkovanje in predlogi pravilnega črkovanja, ki jih dobi uporabnik, so dejansko uporabni.

Query Latency Optimization

Predavanje se je ukvarjalo s časi iskanja, ki je eden najpomembnejših dejavnikov za zadovoljstvo uporabnikov. Razloženo je bilo nekaj dejavnikov, ki so velikokrat vzrok za počasna iskanja in ki se jih da popravit oziroma zaobiti z uporabo drugih metod oziroma z drugačno konfiguracijo iskalnika Lucene.

Reference

[1] http://www.lucenerevolution.org/2013/Lucene-Solr-Revolution-2013-Dublin-Presentations

Stašo Vobič