

Paralelizácia databázových dotazov v aplikácii MedSavant

Miroslav Cupák

Diplomová práca

10. 2. 2014

Problém

- aplikácia: vyhľadávač nad genetickými variáciami (MedSavant)
 - problém: jednovláknové spracovanie dotazov v databáze (Infobright Community Edition)
 - riešenie: paralelizácia spracovania databázových dotazov na úrovni aplikácie (sharding)
 - cieľ: vyšší výkon a škálovateľnosť
-
- spolupráca: Centre for Computational Medicine (SickKids Research Institute) a Computational Biology Lab (University of Toronto)

MedSavant

MedSavant

File View Help

Project Variants Clinic Admin

App Store root

Spreadsheet Search Bar Inspector

Browser

Charts

6.252 (0.5%)
of all variants pass search conditions

Chromosome is chr1
and Position is > 24,361,714
and OMIM is OMIM:ANGELMAN SYNDROME

Type search condition

Search

Type to search page

More Fields

DNA ID	...	Position	V...	...
KB_174_26528	chr1	26.127,020	G	A	61.5	SNP	Hom...	
KB_174_26528	chr1	26.127,202	T	G	176	SNP	Hom...	
KB_174_26528	chr1	26.127,203	T	C	165	SNP	Hom...	
KB_174_26528	chr1	26.127,425	G	A	176	SNP	Hetero	
KB_174_26528	chr1	26.129,438	G	T	30	SNP	Hetero	
KB_174_26528	chr1	26.131,459	G	A	144	SNP	Hetero	
KB_174_26528	chr1	26.131,654	G	A	222	SNP	Hom...	
KB_174_26528	chr1	26.133,099	A	G	62	SNP	Hom...	
KB_174_26528	chr1	26.134,833	T	C	69.5	SNP	Hom...	
KB_174_26528	chr1	26.134,926	C	G	147	SNP	Hetero	
KB_174_26528	chr1	26.135,741	C	G	114	SNP	Hetero	
KB_174_26528	chr1	26.135,913	T	C	84.3	SNP	Hom...	
KB_174_26528	chr1	26.136,452	A	C	222	SNP	Hom...	
KB_174_26528	chr1	26.138,136	C	A	222	SNP	Hom...	
KB_174_26528	chr1	26.138,262	T	C	222	SNP	Hom...	
KB_174_26528	chr1	26.138,451	A	G	71.5	SNP	Hom...	
KB_174_26528	chr1	26.139,055	(too l...	(too l...	176	SNP	Hom...	
KB_174_26528	chr1	26.139,137	C	T	222	SNP	Hom...	
KB_174_26528	chr1	26.139,392	A	G	147	SNP	Hom...	
KB_174_26528	chr1	26.139,444	G	A	111	SNP	Hom...	
KB_174_26528	chr1	26.139,679	C	G	131	SNP	Hom...	
KB_174_26528	chr1	26.139,919	C	T	163	SNP	Hom...	
KB_174_26528	chr1	26.140,573	C	A	222	SNP	Hom...	
KB_174_26528	chr1	26.141,803	G	T	99.5	SNP	Hom...	
KB_174_26528	chr1	27.121,921	C	A	49	SNP	Hetero	
KB_174_26528	chr1	27.124,545	A	G	48	SNP	Hetero	
KB_174_26528	chr1	27.238,150	A	G	78	SNP	Hetero	
KB_174_26528	chr1	33.252,099	(too l...	TCAC...	182	Delet...	Hetero	
KB_174_26528	chr1	33.252,687	T	C	76.5	SNP	Hom...	
KB_174_26528	chr1	33.256,884	C	A	225	SNP	Hetero	
KB_174_26528	chr1	33.275,981	T	C	137	SNP	Hetero	
KB_174_26528	chr1	33.276,424	G	A	222	SNP	Hom...	

Showing 1 - 500 of 6,252

Page 1 of 13

Per page: 500

Variant Gene

Basic Variant Information

POSITION chr1:26,127,202

REFERENCE T

ALTERNATE G

TYPE SNP

GENES SEPNI 18.0 kbp

Detailed Variant Information

DNA ID KB_174_26528

ZYGOSITY HomoAlt

QUALITY 176

DBSNP ID NULL

REF SHOW

Comments

Submit

Individuals with a variant at this position

KB_174_170258

KB_174_26528

KB_174_10-462

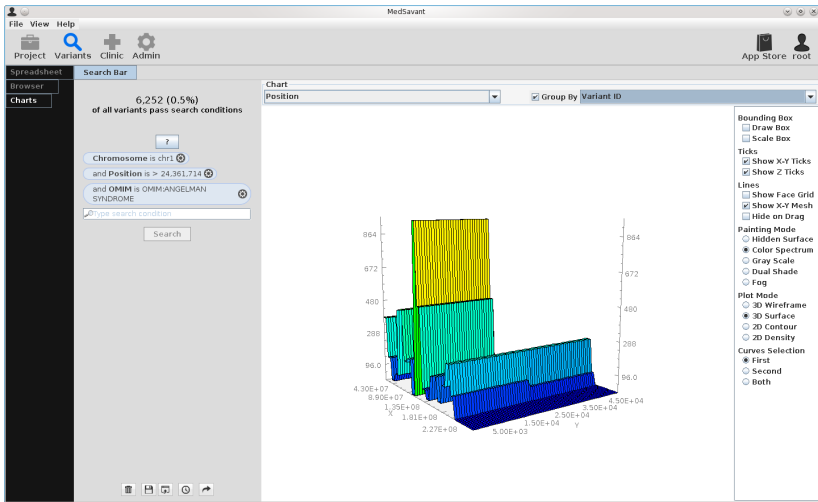
MedSavant

The screenshot displays the MedSavant web application interface. The top navigation bar includes links for File, View, and Help. Below this, there are icons for Project, Variants, Clinic, and Admin. The main content area is divided into several sections:

- Search Bar:** Contains a search bar with the text "6,252 (0.5%) of all variants pass search conditions". Below the search bar, there are filters: "Chromosome is chr1", "and Position is > 24,361,714", and "and OMIM is OMIM:ANGELMAN SYNDROME". A "Search" button is located below the filters.
- Genome Browser:** A visual representation of the human genome with chromosomes 1 through 22, X, and Y. Chromosome 1 is highlighted in blue.
- Location:** A dropdown menu showing "chr1: 26,127,182 - 26,127,223" with a "Go" button and "Length: 42".
- Allele Frequency Table:** A table showing the frequency of alleles at the specified location. The table has columns for Name, Type, Position, Ref, and Alt. The data is as follows:

Name	Type	Position	Ref	Alt
SNP		261272...	T	G
SNP		261272...	T	C
- LD Plot:** A plot showing the linkage disequilibrium (LD) between the variants at the specified location. The plot shows a strong positive correlation between the two variants.
- Gene Model:** A diagram showing the structure of the gene, including exons and introns. The gene is labeled "SEPNI1".
- Filtered Variants:** A list of variants that have been filtered based on the search criteria. The variants are listed with their IDs: KB_174_PHS1-1, KB_174_B1272, KB_174_26528, KB_174_170258, and KB_174_10462. A "Tools" button is located next to the list.

MedSavant



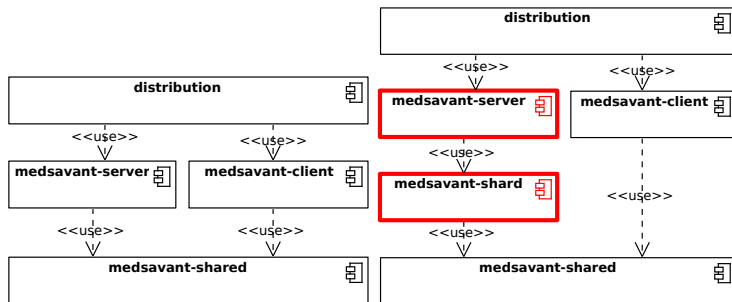
Sharding

- z angl. rozlomenie, rozbitie
- SHAReD-nothING architektúra
- horizontálne rozdelenie množiny dát na nezávislé servery
- distribúcia dotazov a agregácia výsledkov
- sharding stratégia

Požiadavky

- vyšší výkon pri spracovaní pomalých dotazov
- lepšia škálovateľnosť
- plná podpora ICE
- využitie viacerých výpočetných jednotiek zariadenia
- podpora dát rozložených na distribuované servery
- integrácia so serverom
- objektovo orientovaný prístup
- jednoduchá migrácia
- dobre merateľný a reprodukovateľný výkon

Modul



- 2 časti
 - všeobecný sharding rámec
 - sharding logika špecifická pre MedSavant

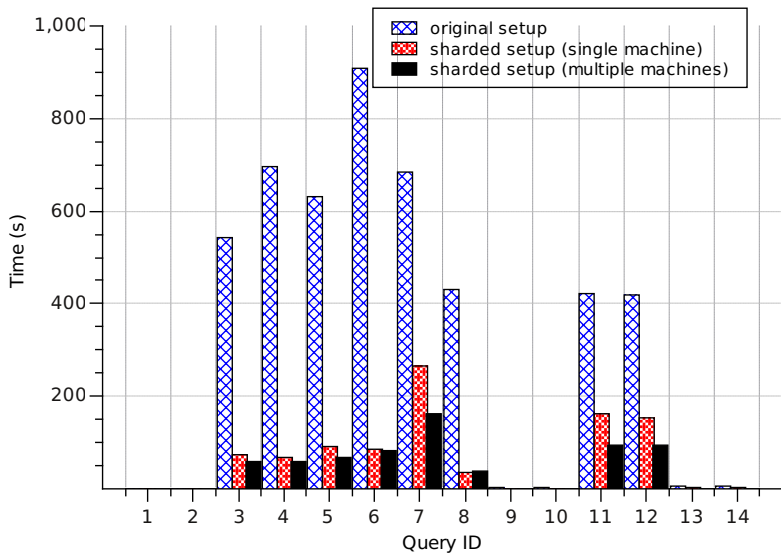
Spracovanie dotazov

- čítanie/zápis
 - rovnaký dotaz na všetky uzly/dotaz pre konkrétny uzol
 - 1 logická databáza/prístup ku konkrétnym uzlom
- fázy spracovania dotazu
 - vygenerovanie pôvodného dotazu v SQL
 - transformácia dotazu do distribuovaného prostredia
 - zaslanie dotazu na databázové servery
 - prijatie čiastkových výsledkov a ich agregácia
 - úprava výsledkov pre potreby aplikácie
- 2 sharding stratégie založené na pozícii v chromozóme
- ďalšie úlohy
 - správa spojení, konfigurácia, udržovanie mapovania...

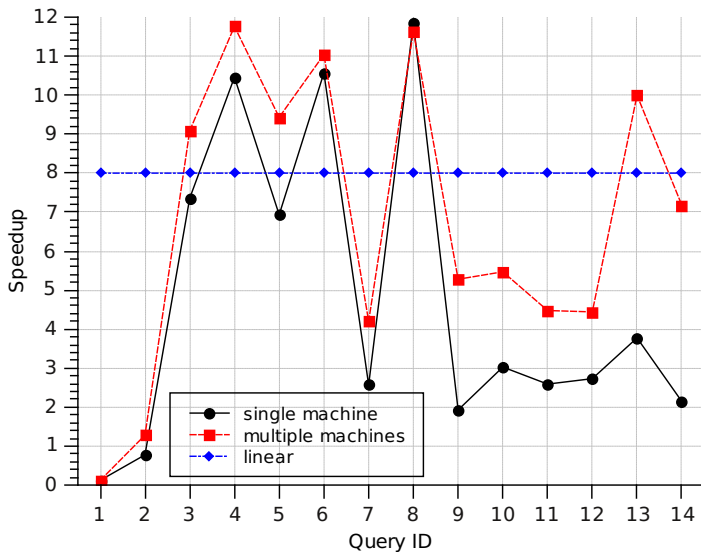
Merania

- 2 aspekty
 - zlepšenie výkonu
 - úspech sharding stratégie
- 2 databázy
 - DB_1 : 134 958 340 variácií
 - DB_2 : 1 378 423 987 variácií
- 2 konfigurácie
 - C_1 : žiadna paralelizácia (1 databáza, pôvodná implementácia)
 - C_2 : paralelizácia v rámci 1 stroja (1 server, 8 shards)
 - C_3 : paralelizácia v distribuovanom prostredí (8 serverov, 8 shards)
- 14 základných dotazov (Q_1 - Q_{14})

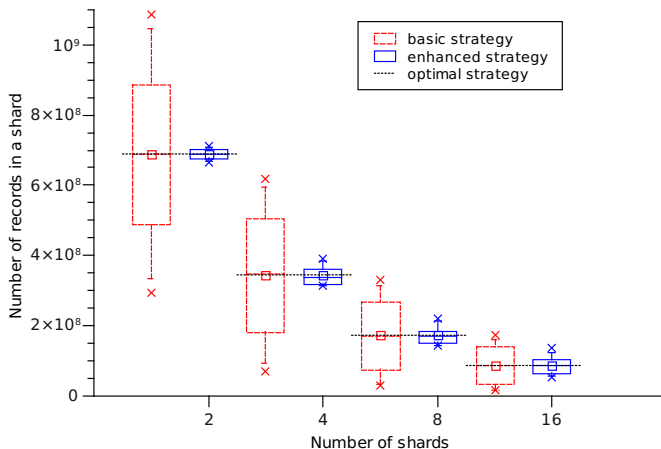
Čas spracovania (DB_2)



Zrýchlenie (DB_2)



Distribúcia dát (DB_2)



- relatívna smerodajná odchýlka **65.44+%**, resp. **4.71+%**

Zhrnutie

- vytvorenie sharding rámca, návrh sharding stratégie pre genetické dáta, plná integrácia s MedSavant
- riešenie spĺňa všetky požiadavky, predovšetkým:
 - vyšší výkon pri spracovaní pomalých dotazov (75+%)
 - dobrá škálovateľnosť (lineárna až superlineárna)
- od odovzdania práce:
 - spätná väzba od CCM
 - demo tento týždeň
 - plány na zaradenie funkcionality do produktu
 - vedecký článok
 - súvisiaci projekt: *sharding in the cloud*

Ďakujem.
Otázky?