groonga

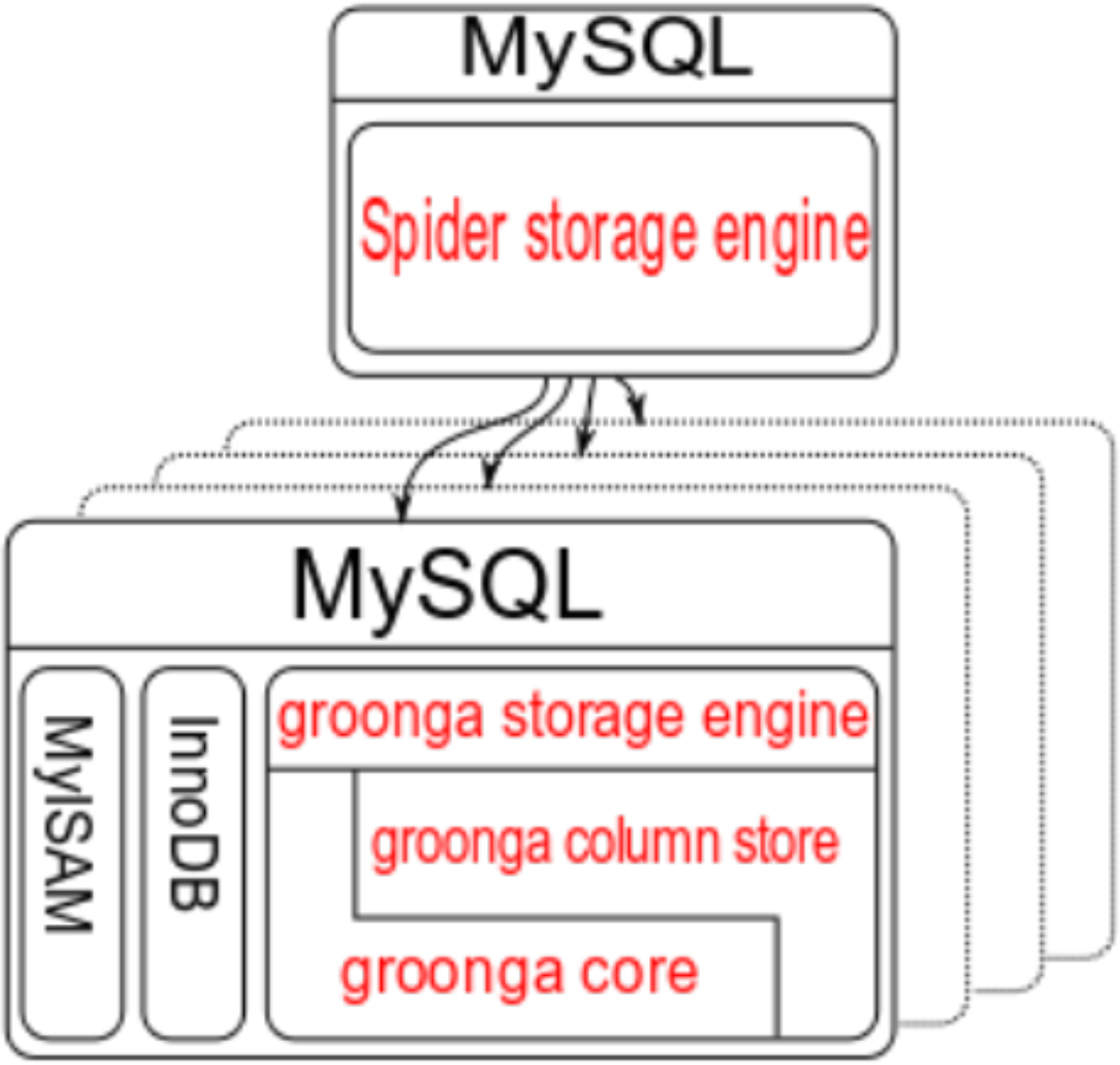
Fulltext Searching
with
Groonga Storage Engine
Brazil Inc.

groonga is...

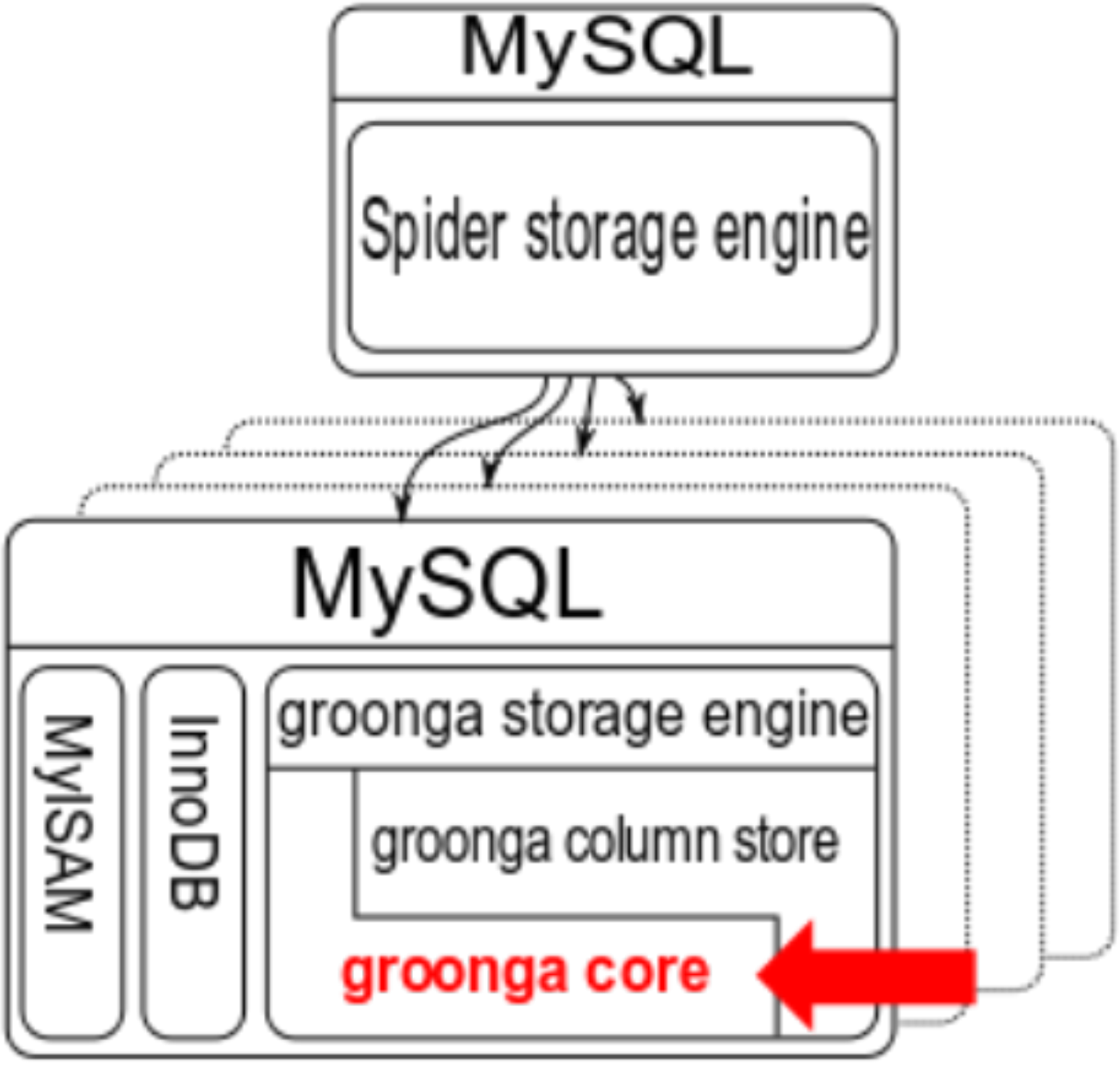
- a Fulltext Search Library
- for Cloud and Web

Characteristics

- Easy to Embed and Scalable
- Highly Precise Search for Any Language
- Fast Searching and Indexing in Realtime

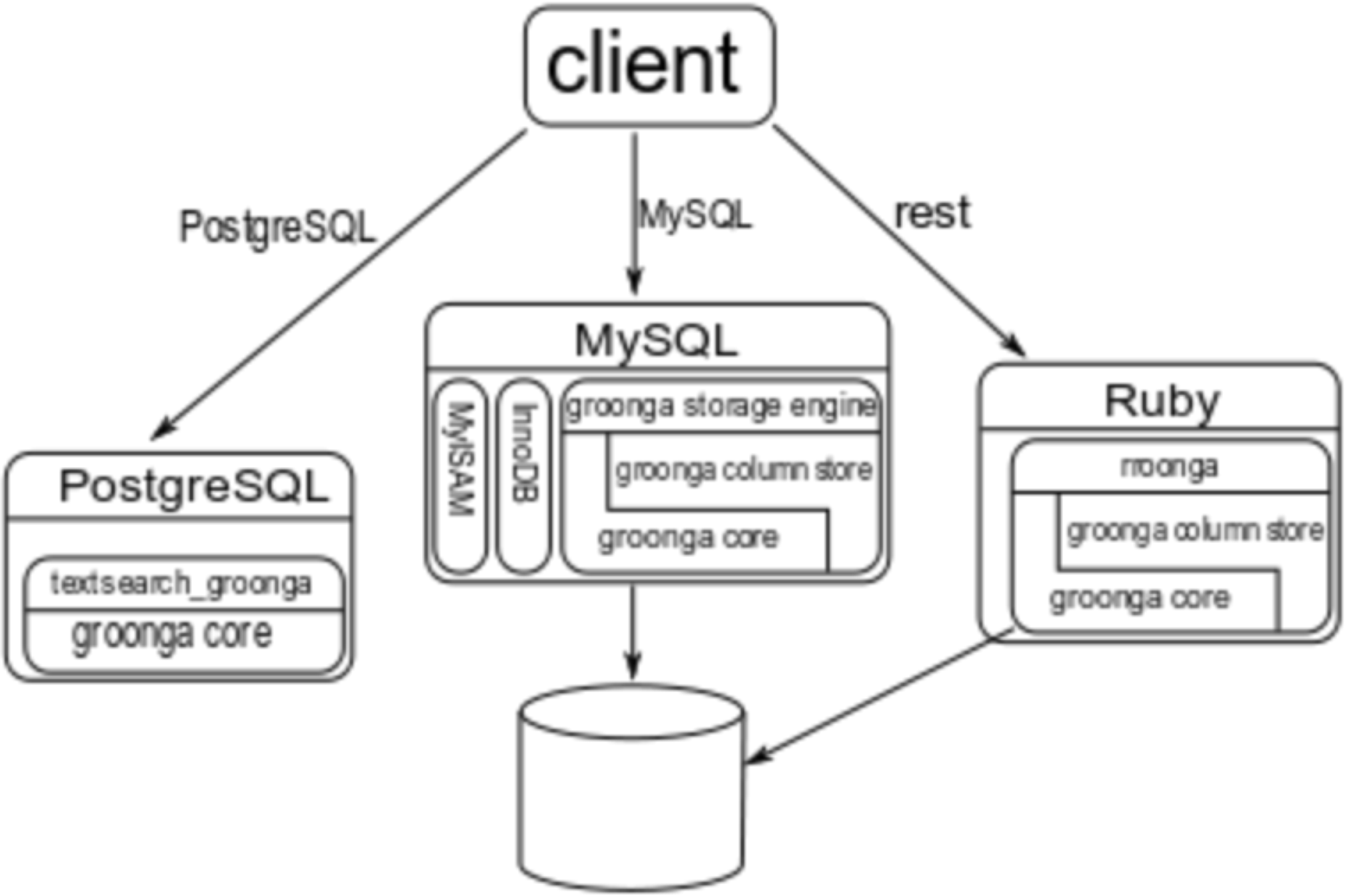

MySQL
Spider storage engine
MySQL
MyISAM
InnoDB
groonga storage engine
groonga column store
groonga core

1. groonga core



Easy to Embed and Scalable

- a Small Library Written in C
- Highly CPU Scalable
- Applicable to Various Process/Thread Model Systems


client
PostgreSQL
MySQL
rest
MySQL
MyISAM
InnoDB
groonga storage engine
groonga column store
groonga core
PostgreSQL
textsearch_groonga
groonga core
Ruby
rroonga
groonga column store
groonga core

How to scale it

- Sharding and Replication is the best and only solution
- for Large Scale Search Engines
- since B.G. era

groonga supports...

- No Built-in Scale-Out Solution
- Suppose to be Embedded Into Various Solutions, like Spider

Highly Precise Search for Any Language

- Importance of Precision
- As for Speed, Money talks
- As for Precision, it can't

Highly Precise Search for Any Language

- Unsegmented Languages need
- not only Pluggable Tokenizers
- but also Solutions for Vague Boundaries Problem

Vague Boundaries Problem

ここではきものをぬいでください

ここで / はきものを / ぬいで / ください。
Put off your shoes here.

ここでは / きものを / ぬいで / ください。
Put off your clothes here.

groonga supports...

Full Inverted Index

- ✓ inevitable for unsegmented languages

Highly Compressed Index

- ✓ no stop words needed

Patricia TRIE Lexicon

- ✓ Partial string match on lexicon

Fast Searching and Indexing in Realtime

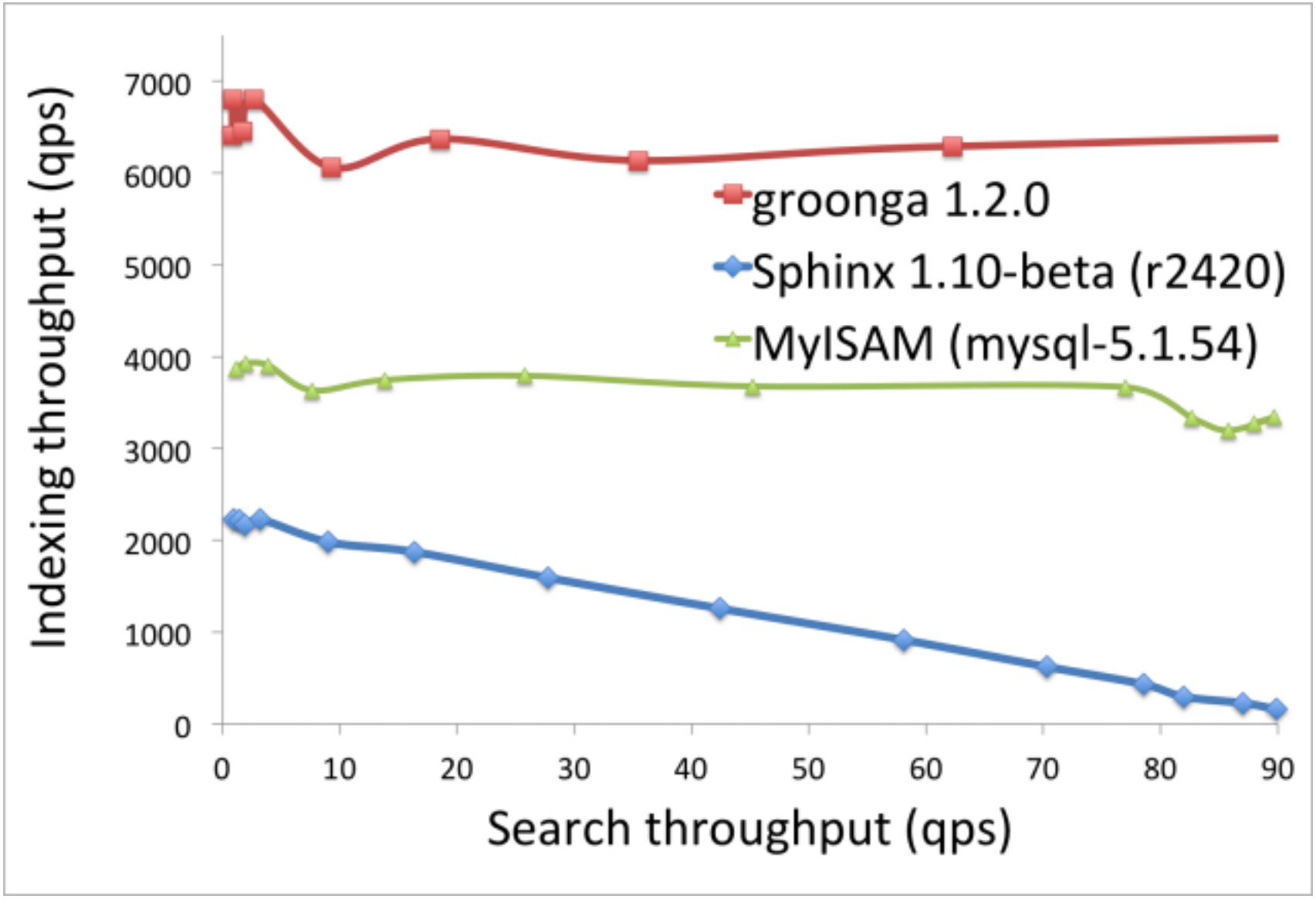
- Static Indexing was Mainstream in Full Inverted Index
- Dynamic(Realtime) Indexing is in Great Demand for Web

Fast Searching and Indexing in Realtime

- Web is Growing Exponentially Every Moment
- And Millions of Users Need to Search from Among Them
- Searching and Indexing must be Performed Simultaneously

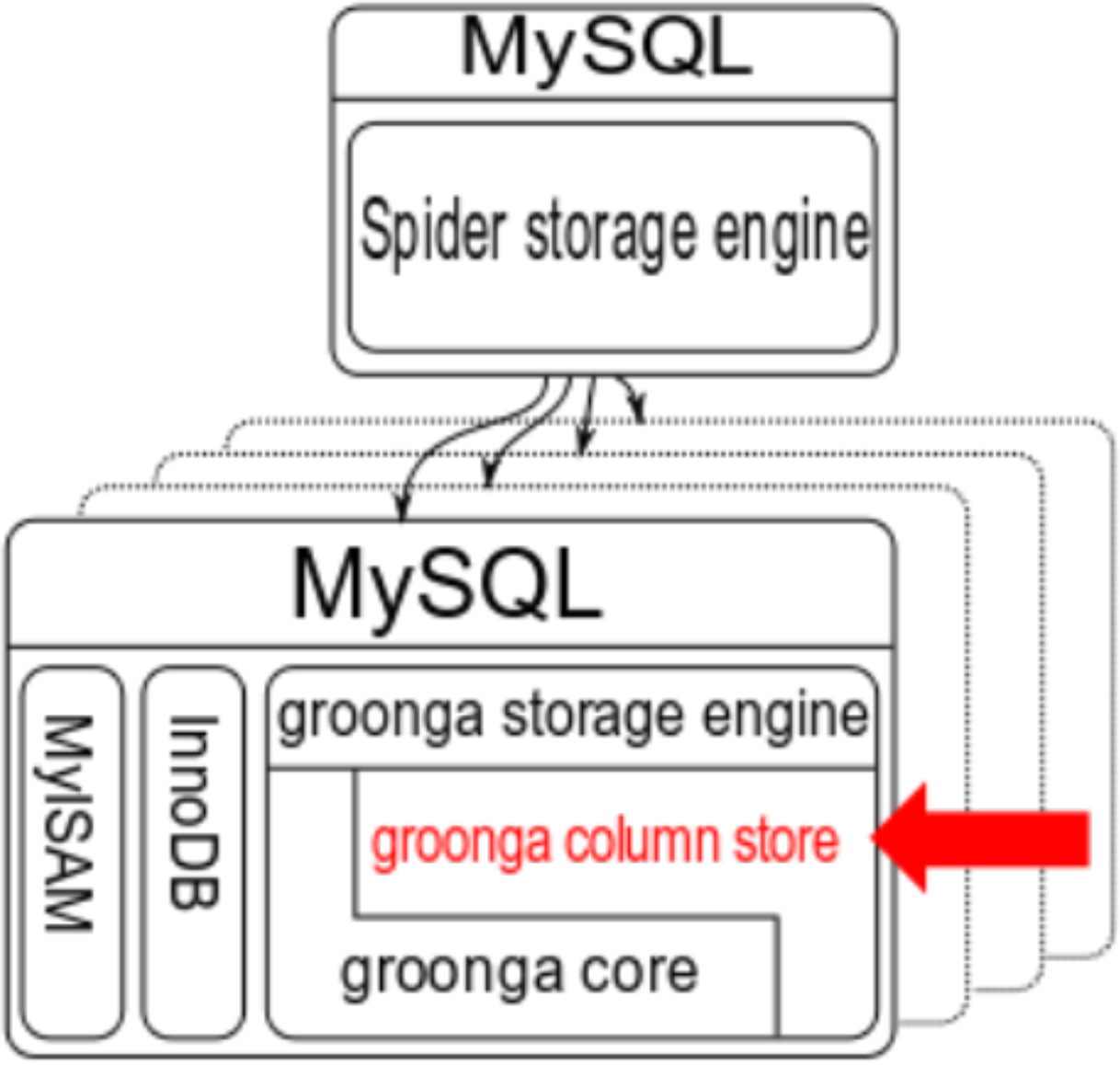
groonga supports...

- Lock Free Index Update
- Indexing Never Block Searching



Performance of Simultaneous Indexing and Searching

2. groonga column store



groonga column store

- Why Own Data Store?
- No Storage Engines were available
- Suitable for Typical Search Engine Queries

Typical Search Engine Query

- hits large number of records
- filtered by multiple conditions
- group by specific conditions
- order by a dynamic condition
- output limited number of records

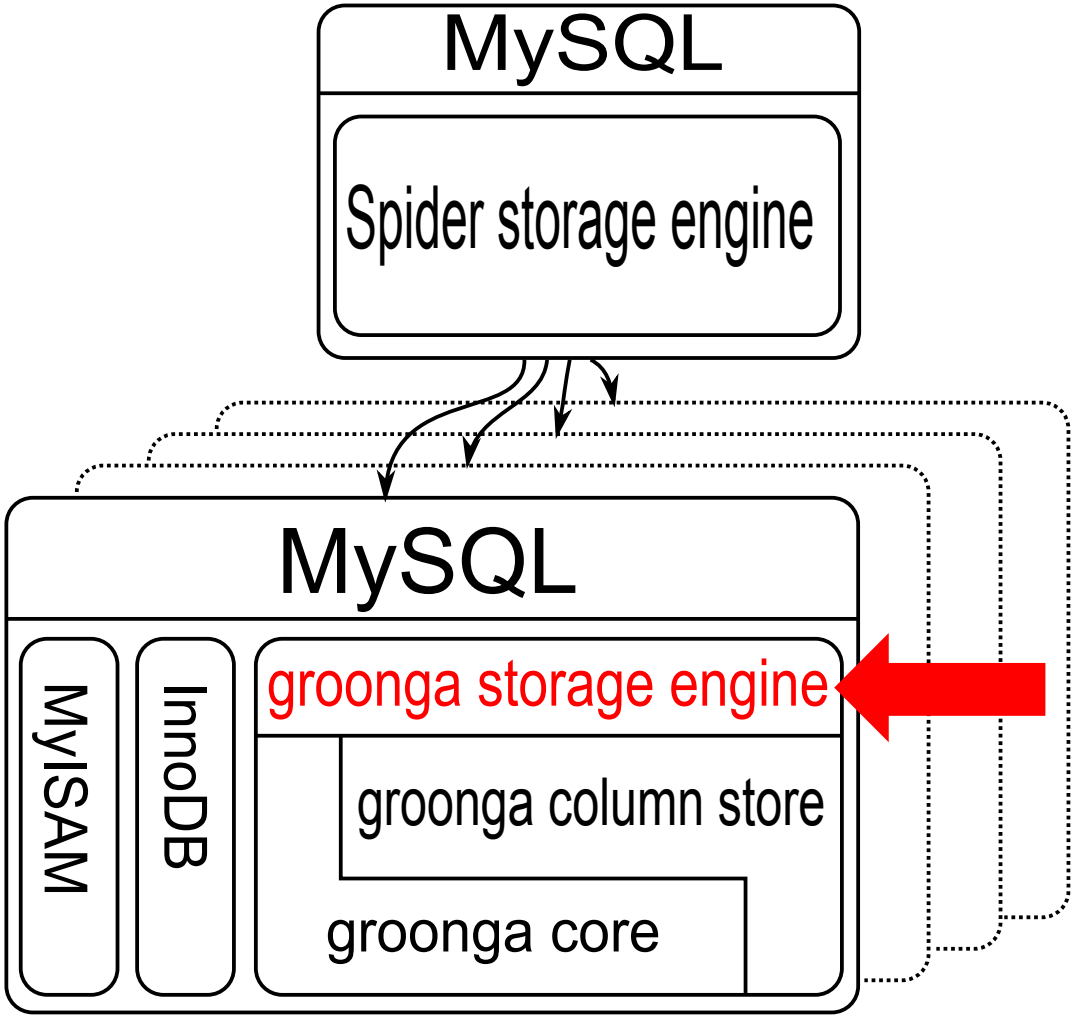
Groonga Column Store supports...

- Reduce I/O Cost of Performing Filter/Sort/Group Operations
- Especially the Number of Hits Become Very Large
- Fast Range Search and Geo Search are also Available

groonga storage engine

Brazil, Inc.
Tasuku SUENAGA
a.k.a. gunyarakun

What I talk about



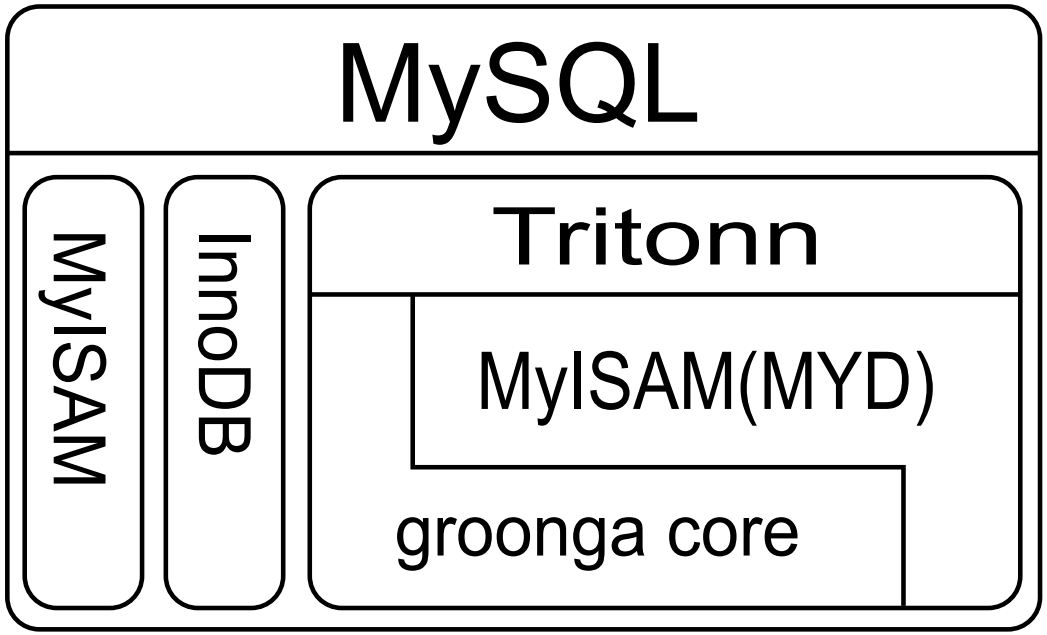
MySQL fulltext index

- Phrase search is slow.
- Updating index is slow.
- Cannot combine full text search index with other indexes(like B-Tree).

Our prior product Tritonn solves.

Tritonn

- Tritonn =
  groonga +
  patches for MyISAM



MySQL v.s. Tritonn

Target dataset : Wikipedia English 458,713 record 1088MB

	MySQL(5.0)'s Fulltext Index	Tritonn
Index size	109 MB	1028 MB
Phrase search for 'united states'	44.91 sec	0.40 sec
Indexing after inserting recs	1,474 sec	1,808 sec
Inserting recs after idx. creation	28,182 sec	1,839 sec
Where MATCH AGAINST and order by primary key	20.33 sec	0.89 sec
Where MATCH AGAINST and primary key > 200000	6.55 sec	0.32 sec

So Tritonn provides …

- Fast phrase search
- Fast index update (realtime)
- Works well with other indexes.

But some problems remain.

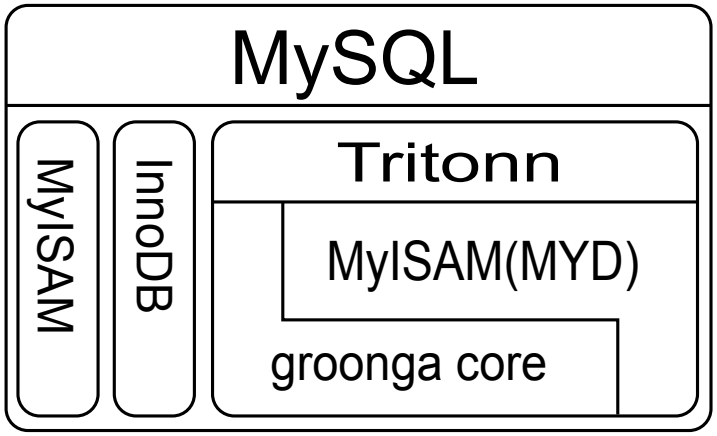
Remaining problems

- MyISAM based
  - Table lock
    - when updating table, read accesses are blocked.
- Patch based
  - Patch maintainance and building patched MySQL is too messy.

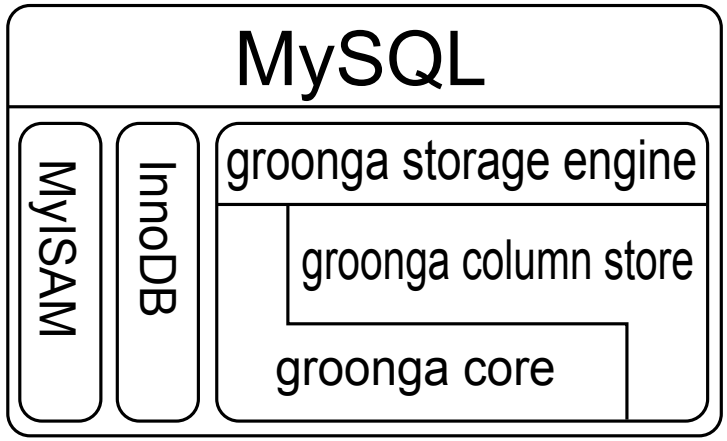
Need for a new solution.

New solution is

- groonga storage engine
  - Use column store of groonga instead of MyISAM.
  - Not patch but storage engine.



Tritonn (old)



groonga storage engine(new)

Advantage

- Table lock free
  - Column store of groonga is lock-free.
- Only access columns required
  - Not row-based.
- Easy to build and develop

And some optimization
for typical queries

Optimization(1)

- COUNT(*) optimization.
  - For queries like below.

```
SELECT COUNT(*) FROM table
WHERE MATCH(col)
AGAINST ('query');
```

Optimization(2)

- ORDER BY score and LIMIT optimization.
  - For queries like below.

```
SELECT * FROM table
WHERE MATCH(col)
AGAINST ('query')
ORDER BY MATCH(col)
AGAINST ('query')
LIMIT 10;
```

Conclusion of my part

- groonga storage engine provides

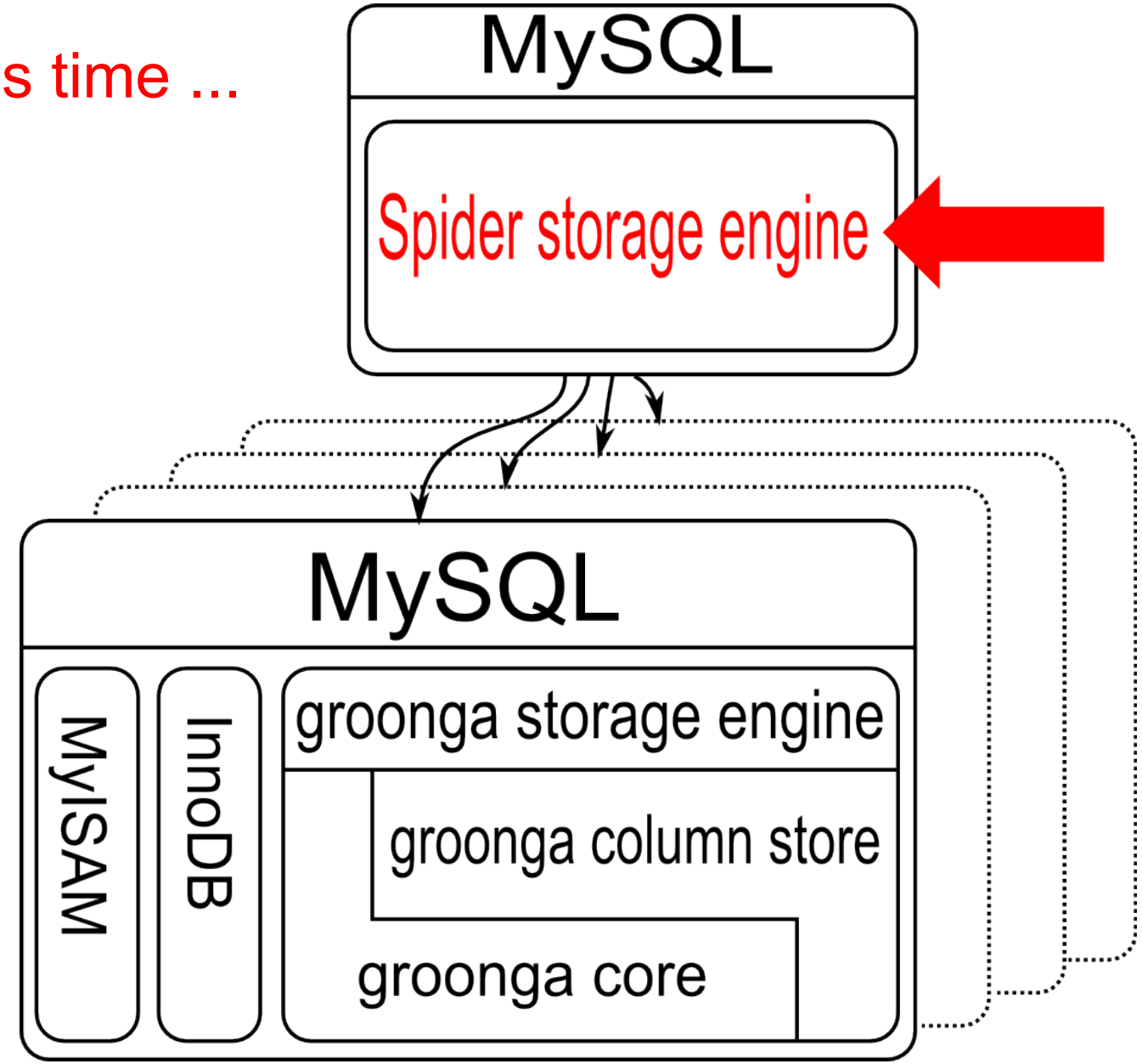
- Fast phrase search
- Fast index update (realtime)
- Inserting records doesn't block reading records

The combination of Groonga and Spider

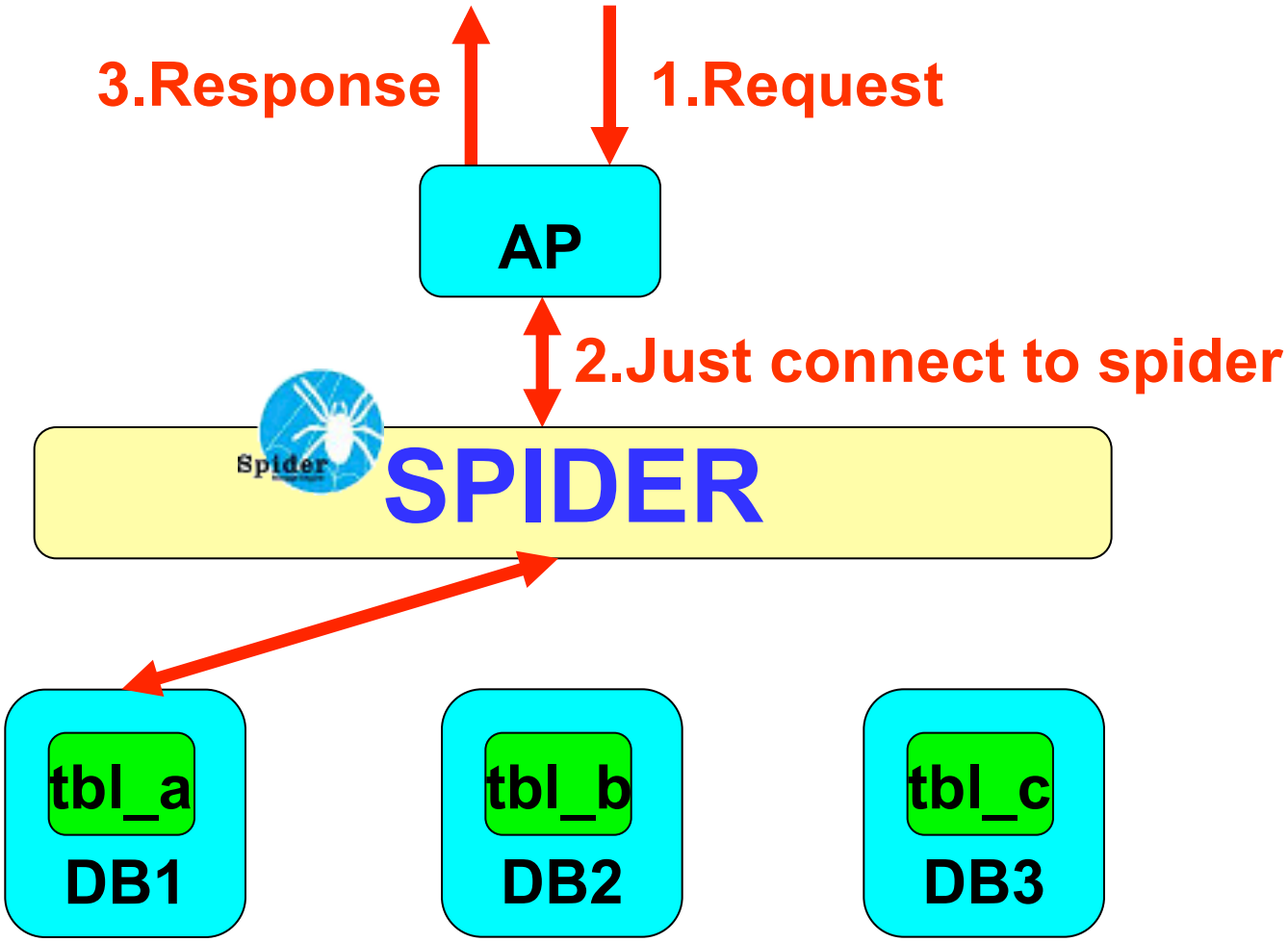
Kentoku SHIBA
kentokushiba at gmail dot com

The combination of Groonga and Spider

In this time ...



What is Spider Storage Engine



Spider Storage Engine is a storage engine for database sharding transparently.

The combination of Groonga and Spider

You can get following power by combination of Groonga and Spider.

- The optimization for the fulltext searching with sorting by score.

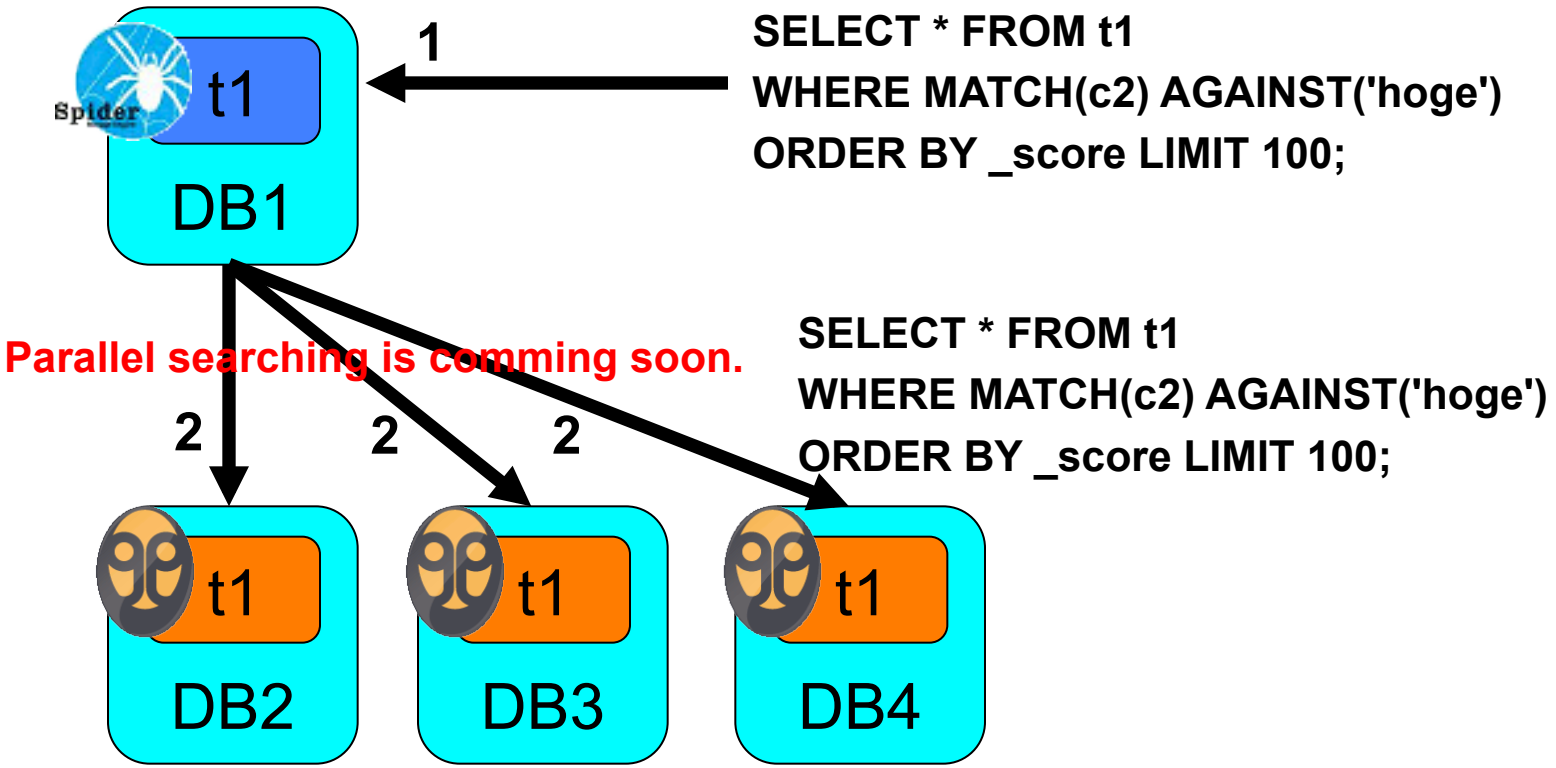
- The optimization for the sorting by range partition key column.

- The optimization for the fulltext searching with filtering by partition key column.

The optimization for the fulltext searching with sorting by score

(The case of scanning all partitions)

Sorting by score

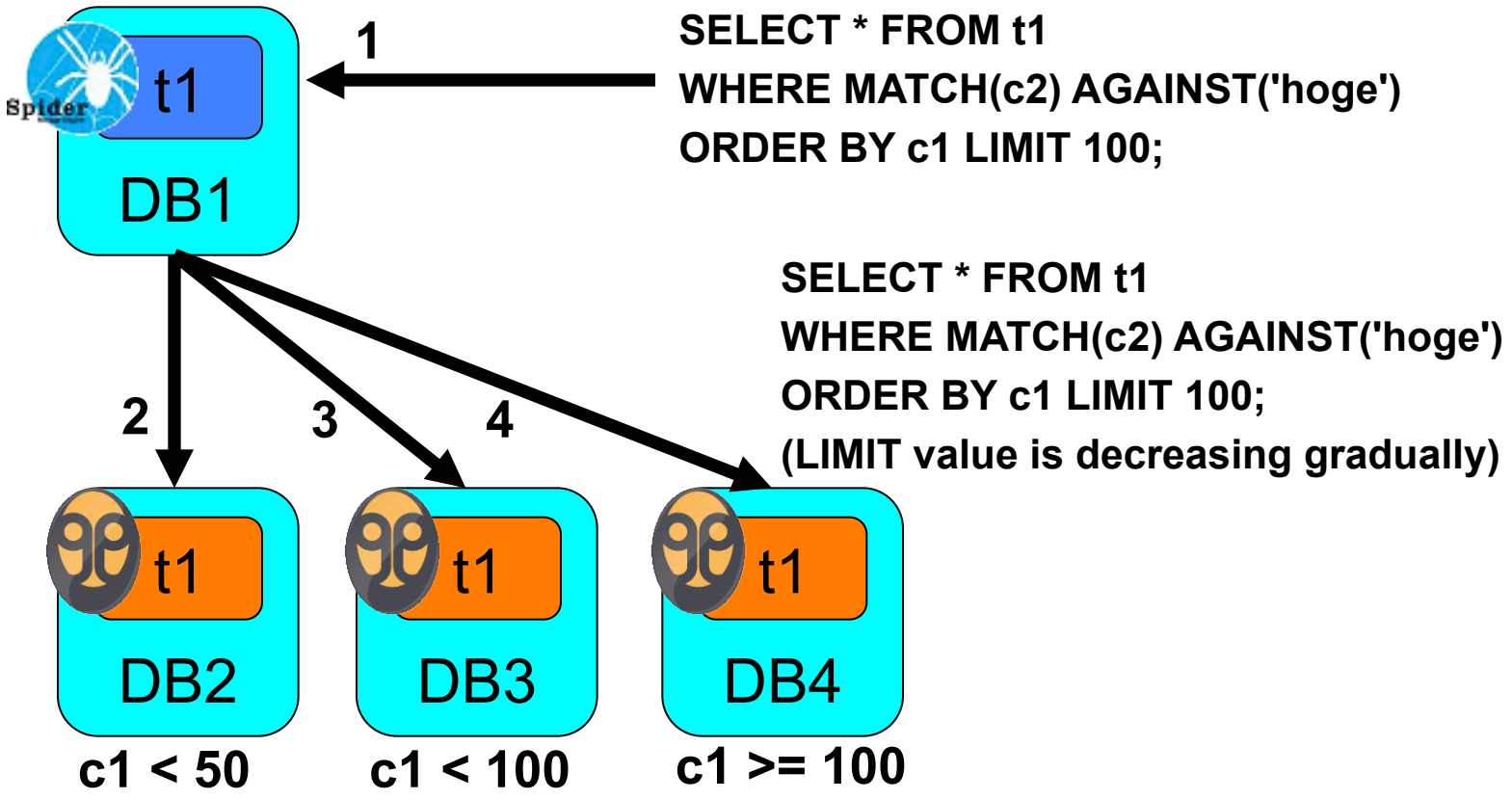


- Parallel searching
- 2 step limitation

The optimization for the sorting by range partition key column

(coming soon)

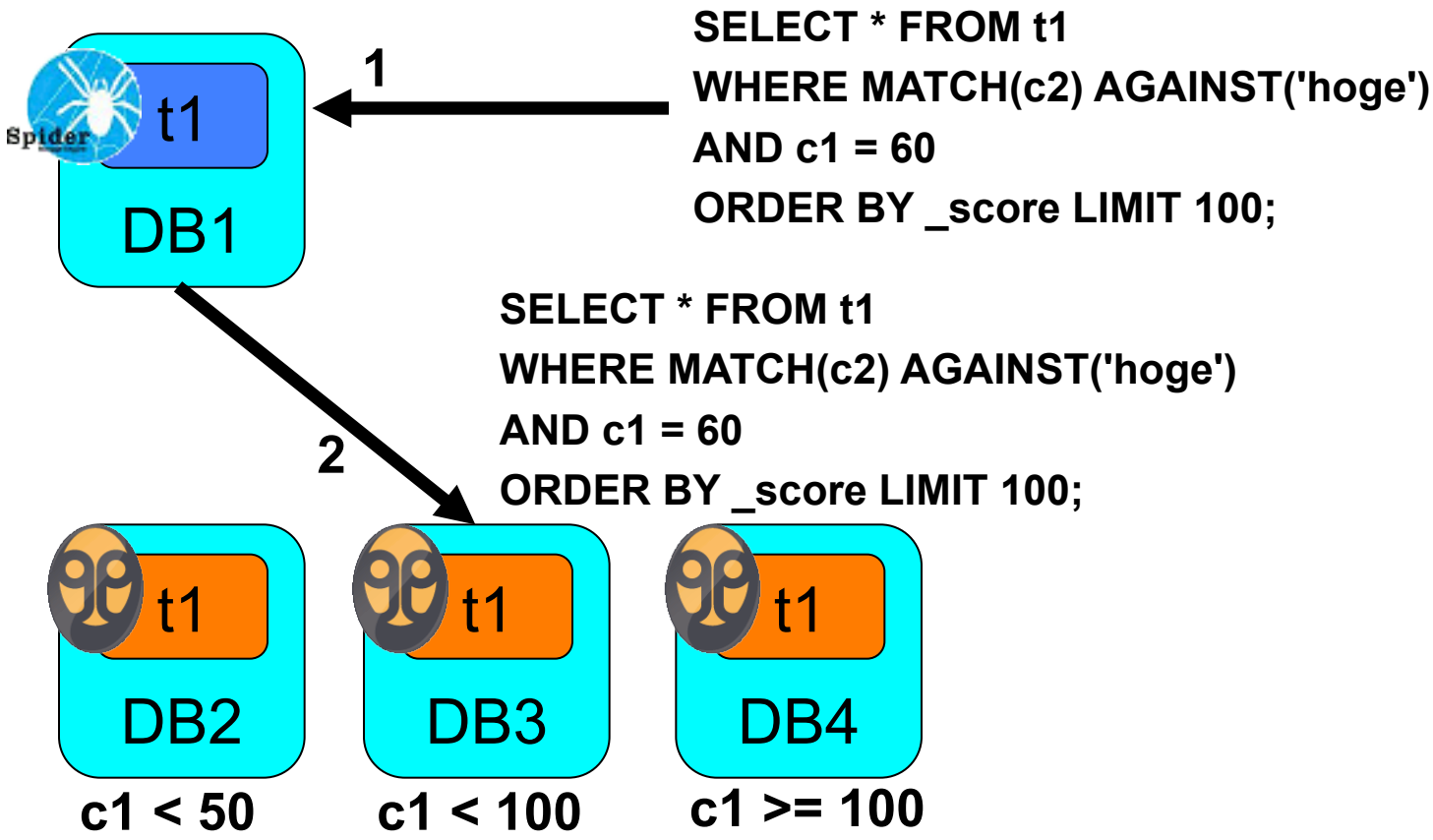
The sorting by range partition key column



- Sort optimization with range partition

The optimization for the fulltext searching with filtering by partition key column

The filtering by partition key column



- Partition pruning

End of the session

Source code and binary

If you want to try introduced Spider features,

you can download from here and try.

source code

http://groonga.org/pkg/mysql-5.5.8-spider-2.24h-vp-0.13-hs-1.0.src.tgz

binary (Linux x86_64 glibc2.3)

http://groonga.org/pkg/mysql-5.5.8-spider-2.24h-vp-0.13-hs-1.0.bin.tgz

initialize SQL

http://groonga.org/pkg/spider-init-2.24-for-5.5.8.tgz

Contact us

If you have some questions, comments or suggestions, please contact us from here.

http://bit.ly/fSs5vx

Thank you for taking your time!!

Daijiro MORI (morita at razil dot jp)
Tasuku SUENAGA (a at razil dot jp)
Kentoku SHIBA (kentokushiba at gmail dot com)