Search query assistance: Autosuggestion

Vitalii Melnychuk, Software Engineer

Intro

10 members

50 servers

6 TB of data

Mysql

Elasticsearch

NODEJS

Docker

AWS

Terraform

KIbana

Jenkins

Grafana

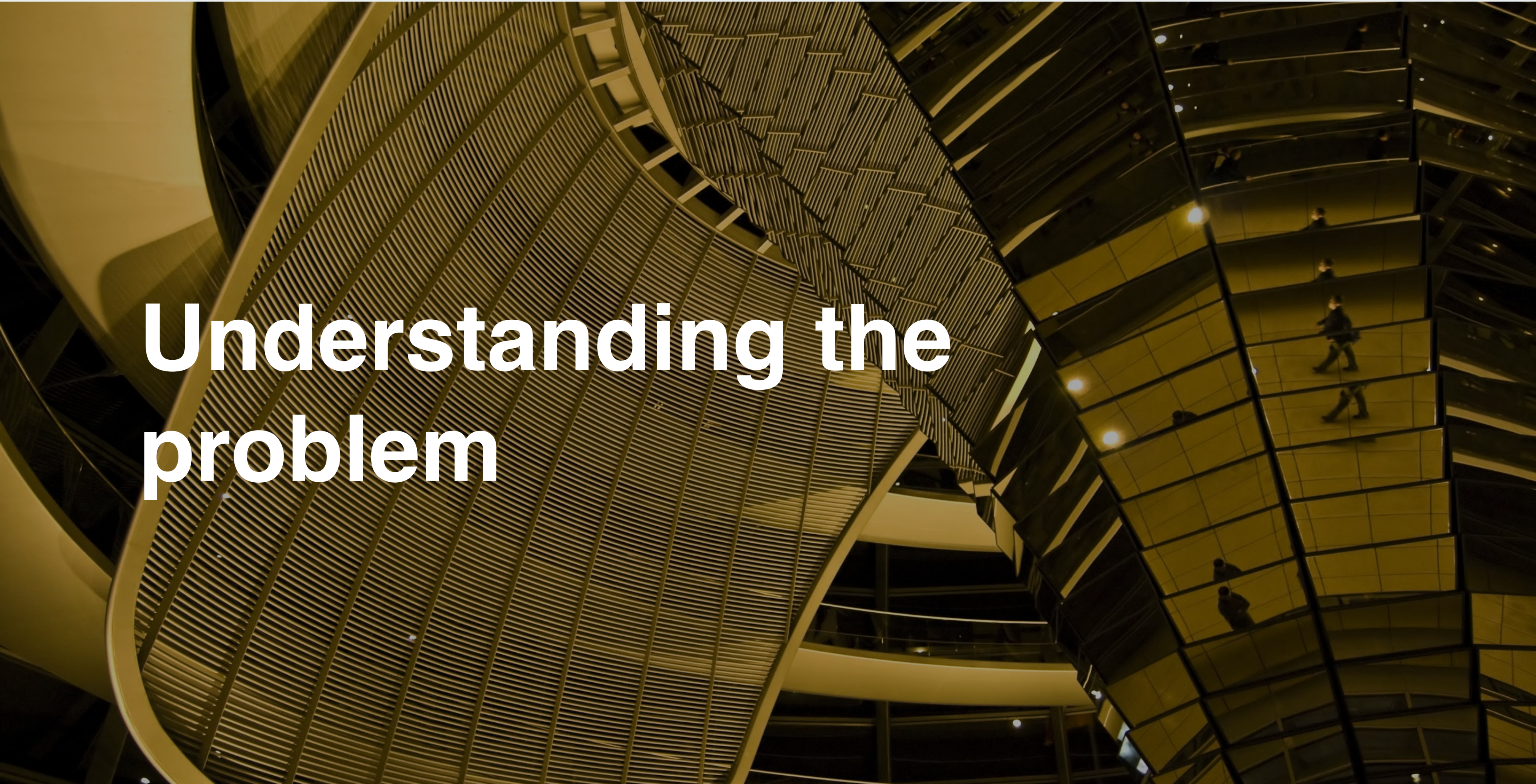
Highload Project - highload solution

1 Highload begins when one physical server becomes unable to handle data processing.

2 If a single instance serves 10,000 connections simultaneously - it's highload.

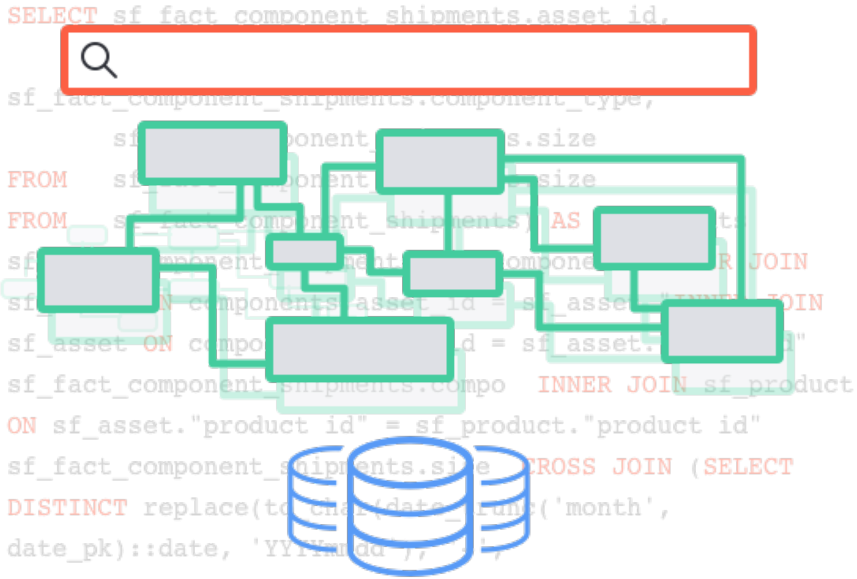
3 Your project is highload if it processes 100+ dynamic requests per second.

4 Usage of Lambda Architecture and Kafka makes the system highload.



Understanding the problem

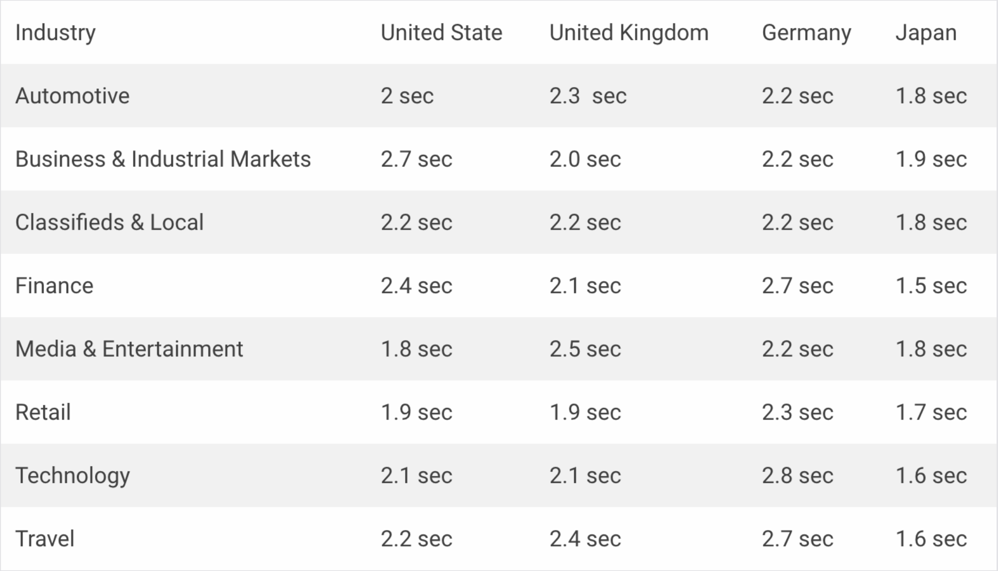
Expectations?


SELECT sf_fact_component_shipments.asset_id,
sf_fact_component_shipments.component_type,
FROM
FROM
INNER JOIN
ON sf_asset."product id" = sf_product."product id"
CROSS JOIN (SELECT
DISTINCT replace(to_char(date_trunc('month',
date_pk)::date, 'YYYYmmdd'),

WISE ENGINEERING

How Fast Should A Website Load?

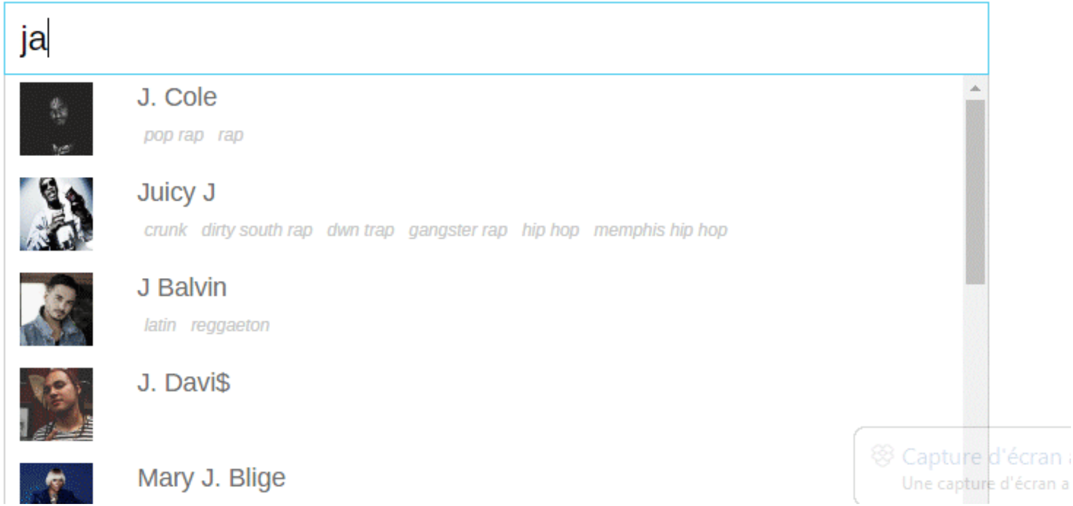
QUOTE: “I wouldn’t worry about it too much. Make it as fast as you reasonably can.” Gary Illyes, Google 2016



Industry	United State	United Kingdom	Germany	Japan
Automotive	2 sec	2.3 sec	2.2 sec	1.8 sec
Business & Industrial Markets	2.7 sec	2.0 sec	2.2 sec	1.9 sec
Classifieds & Local	2.2 sec	2.2 sec	2.2 sec	1.8 sec
Finance	2.4 sec	2.1 sec	2.7 sec	1.5 sec
Media & Entertainment	1.8 sec	2.5 sec	2.2 sec	1.8 sec
Retail	1.9 sec	1.9 sec	2.3 sec	1.7 sec
Technology	2.1 sec	2.1 sec	2.8 sec	1.6 sec
Travel	2.2 sec	2.4 sec	2.7 sec	1.6 sec

WISE ENGINEERING

What we have to build?



ja

J. Cole

pop rap rap

Juicy J

crunk dirty south rap dwn trap gangster rap hip hop memphis hip hop

J Balvin

latin reggaeton

J. Davi$

Mary J. Blige

Capture d'écran

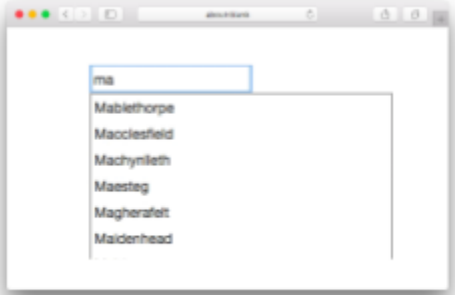
Une capture d'écran a

Solutions analysis

- Frontend side filter
- Mysql `LIKE` search
- Elasticsearch Completion Suggester
- Own Data structures

Data

Code



ma

Mablethorpe

Macclesfield

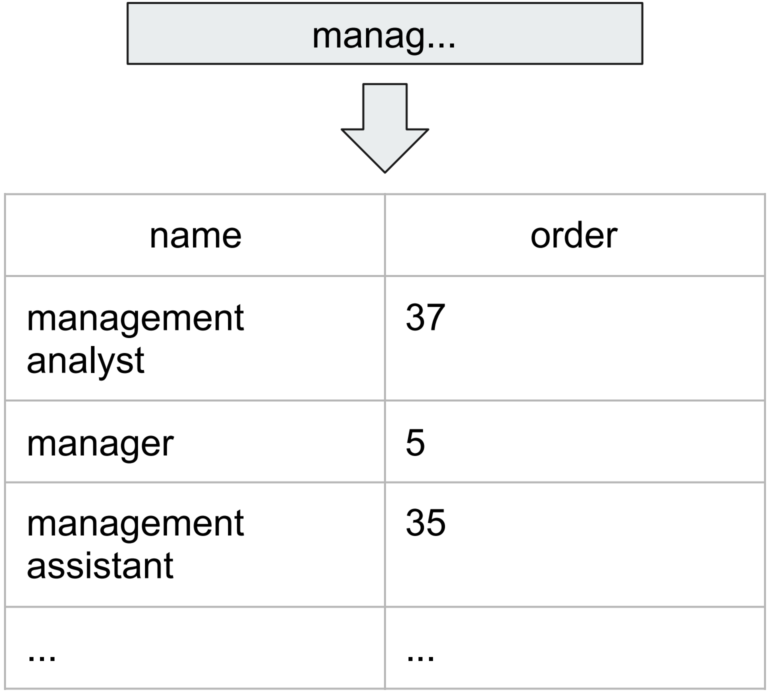
Machynlleth

Maesteg

Magherafelt

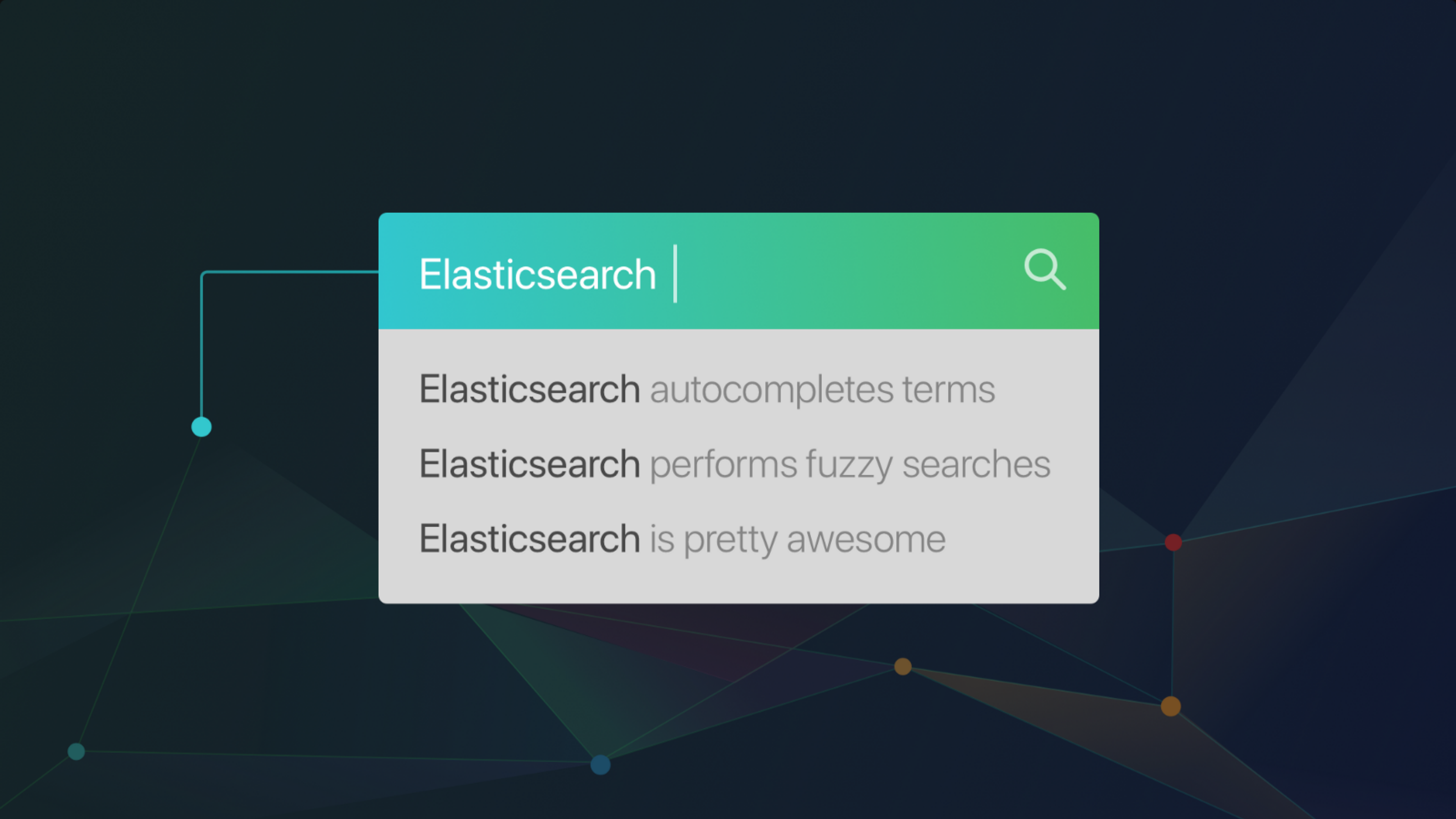
Maidenhead

Front end



manag...

name	order
management analyst	37
manager	5
management assistant	35
...	...



Elasticsearch

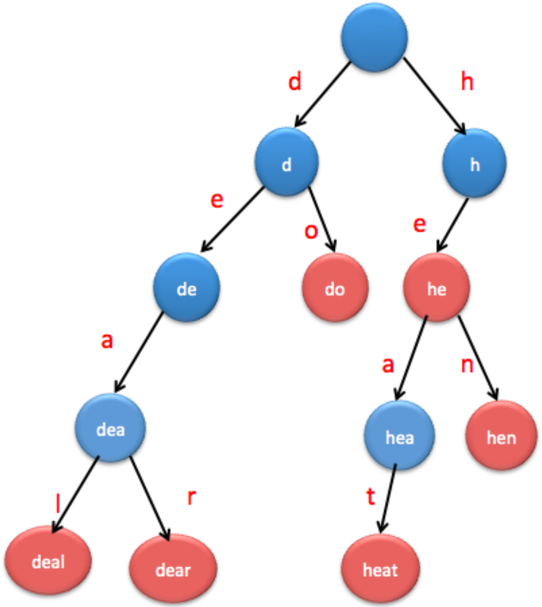
Elasticsearch autocompletes terms

Elasticsearch performs fuzzy searches

Elasticsearch is pretty awesome

Trie Structure

1. With Trie, we can insert and find strings in ***O(L)*** time where *L* represent the length of a single word.
2. We can easily print all words in alphabetical order which is not easily possible with hashing.
3. We can efficiently do prefix search with Trie.



WISE ENGINEERING

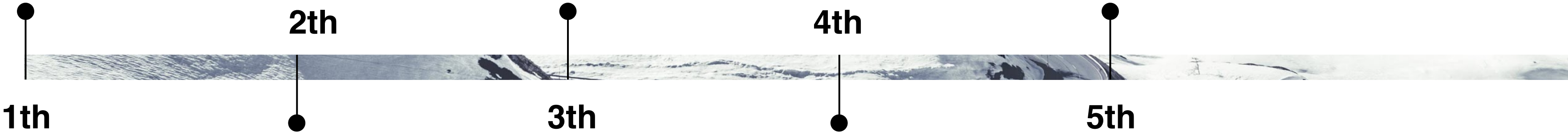
Performance platform

- 100 threads
- 50 req/sec
- 50 000 samples
- c3.large

Performance tests

	Mysql	Elasticsearch	In-memory
Average response	100ms	130ms	10ms
Throughput	~650 req/sec	500 req/sec	~1100 req/sec
Errors %	2%	5%	0%

Vision



1th

Build data before application started

2th

Trie sorted by alphabet

3th

Prioritize keywords

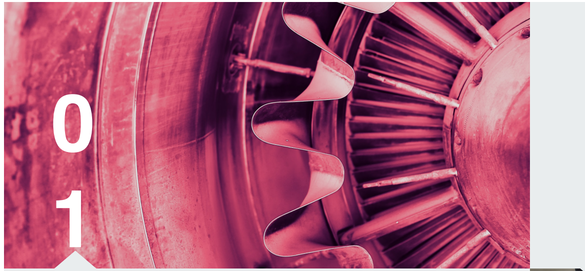
4th

Store only addresses

5th

Memory-optimized structure

Process



Download all keywords

Build a trie

Respond by using in-memory trie structure

WISE ENGINEERING

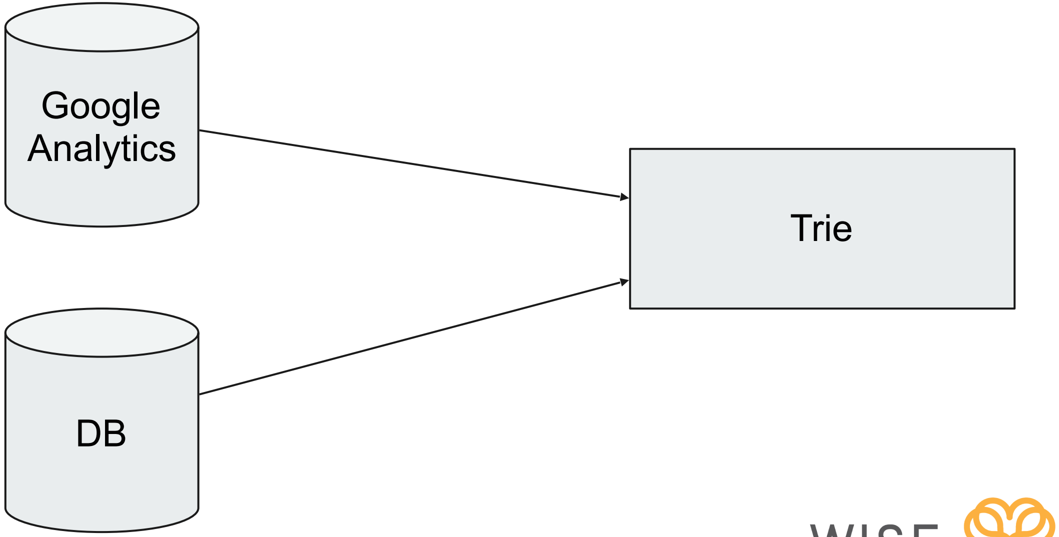
Disadvantages:

- Self-managed solution
- Cannot analyze full text
- Complicated in implementation

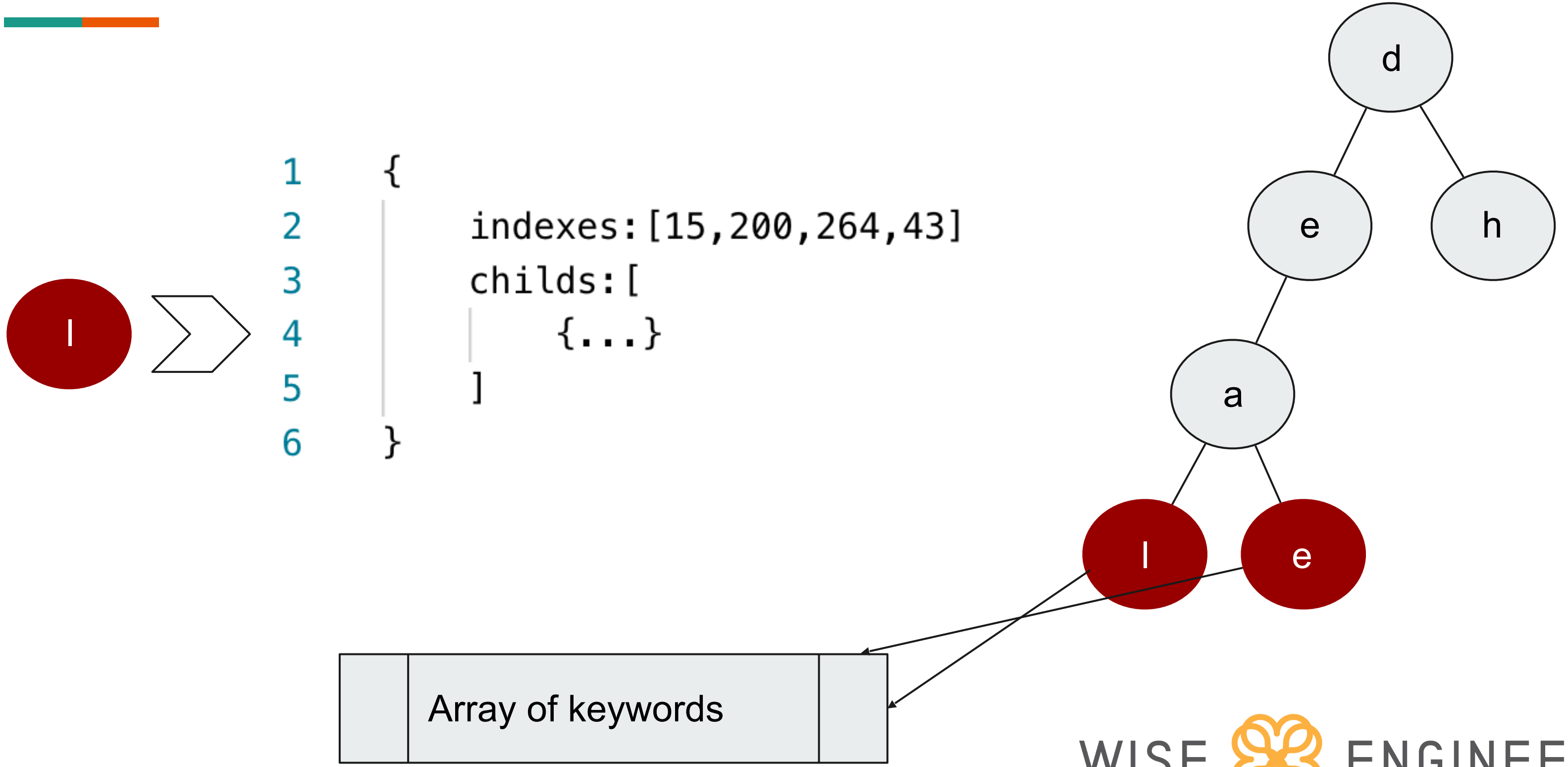
: Advantages

- Scalable and high-load search engine
- response time

Not alphabetical search



Make some modifications



```
{
    indexes:[15,200,264,43]
    childs:[
        {...}
    ]
}
```

https://github.com/**melnychukvitaliy/trie**

Proposed solution

You can save time in the place that doesn’t affect user experience and respond quickly.

Results

50K

Keywords in our mysql databases sorted by priority

25mb

Memory used in order to build trie structure

<10ms

Response

Thank you

WISE ENGINEERING

Questions?