# Apache Lucene/Solr Internals

# About me

Java and all around

Principal Software Engineer
at Grid Dynamics

Kharkiv

asokolenko@griddynamics.com


Grid Dynamics
Scalable eCommerce Platform Solutions

the magic of macy's .com

# Apache Lucene/Solr Internals

# stackoverflow

Lucene.net

June 2013 database

14.630.209 records

VM
16 CPU cores
16 GB memory

4 nodes
× 12GB disk space

Indexing took 5 hours
in 100 threads
1000 batch

lightweight
performant
search
library



elasticsearch.

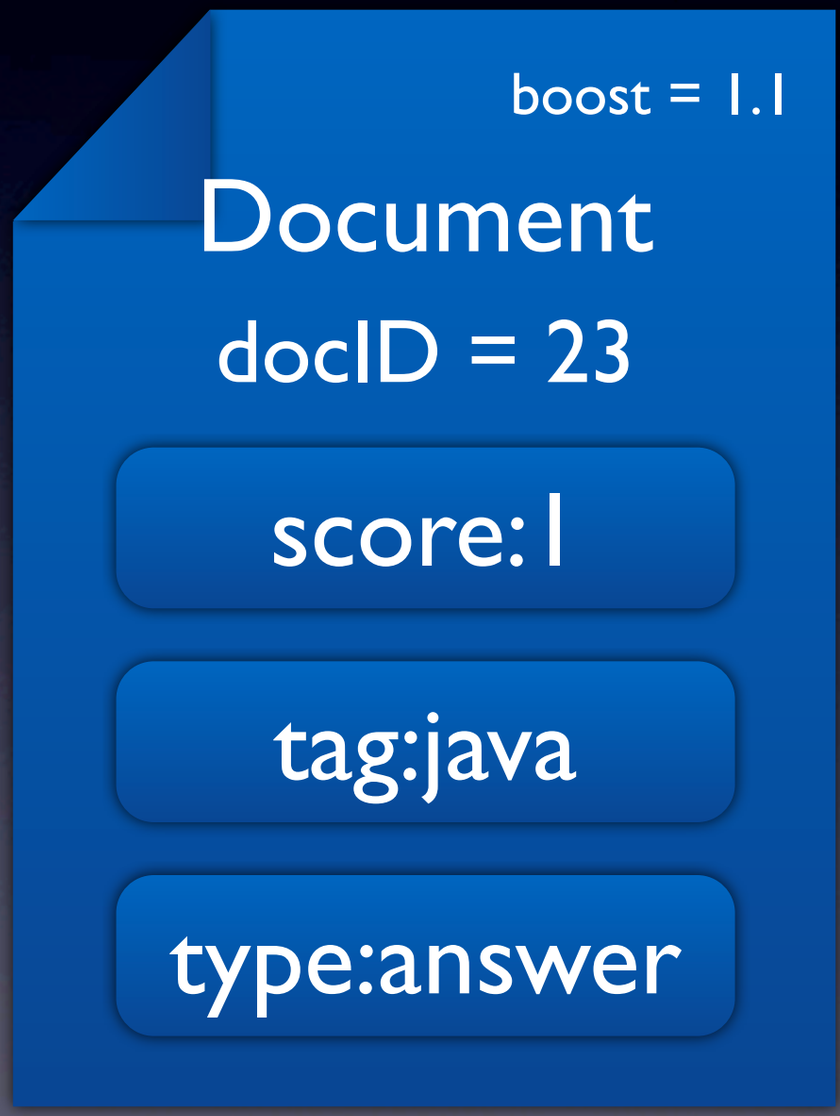

Apache Solr

# Data Model

- document oriented
- flat
- store
- index

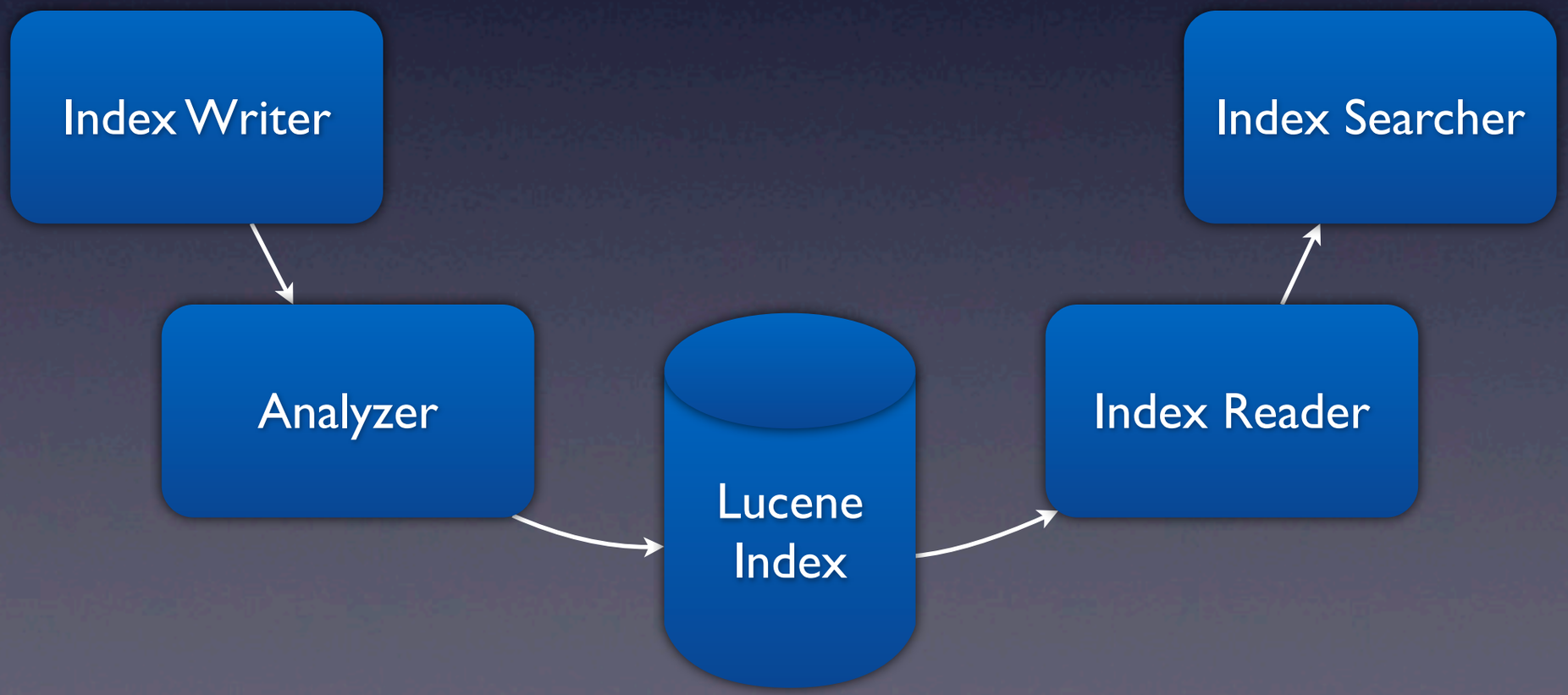
# Data Model

- document oriented
- flat
- store
- index

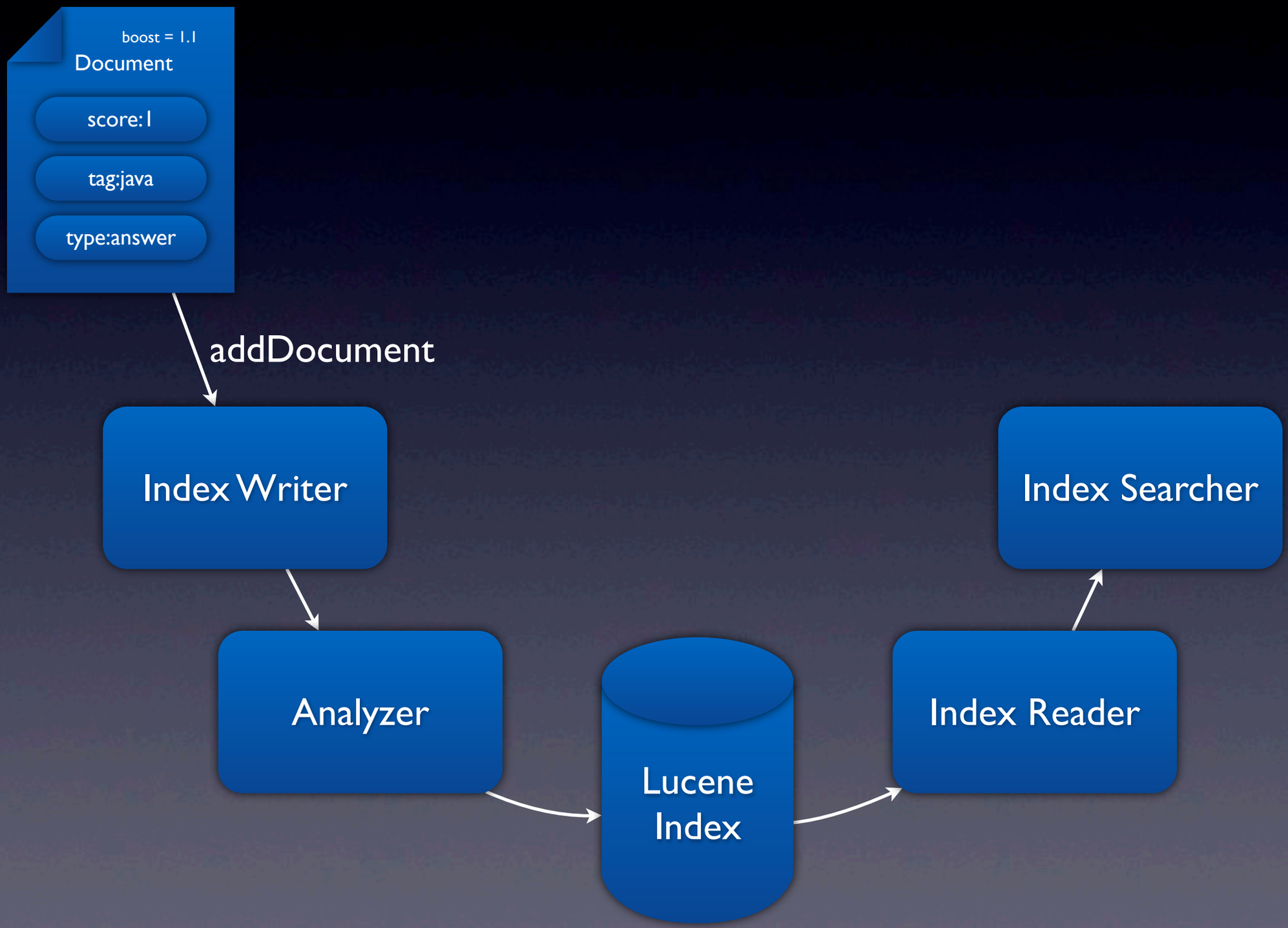


# Showcase

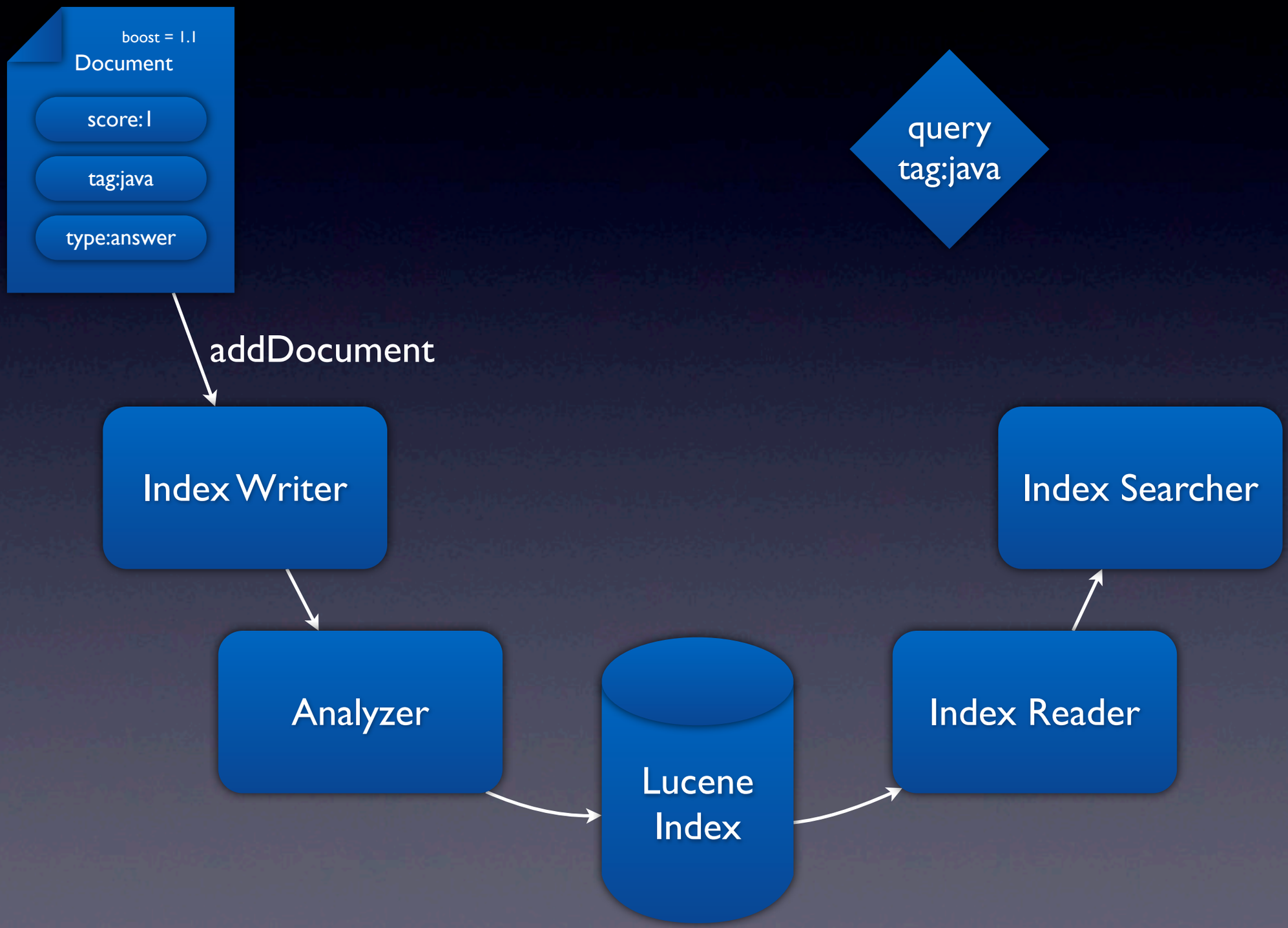
# Basic Flow


Index Writer
Analyzer
Lucene Index
Index Reader
Index Searcher

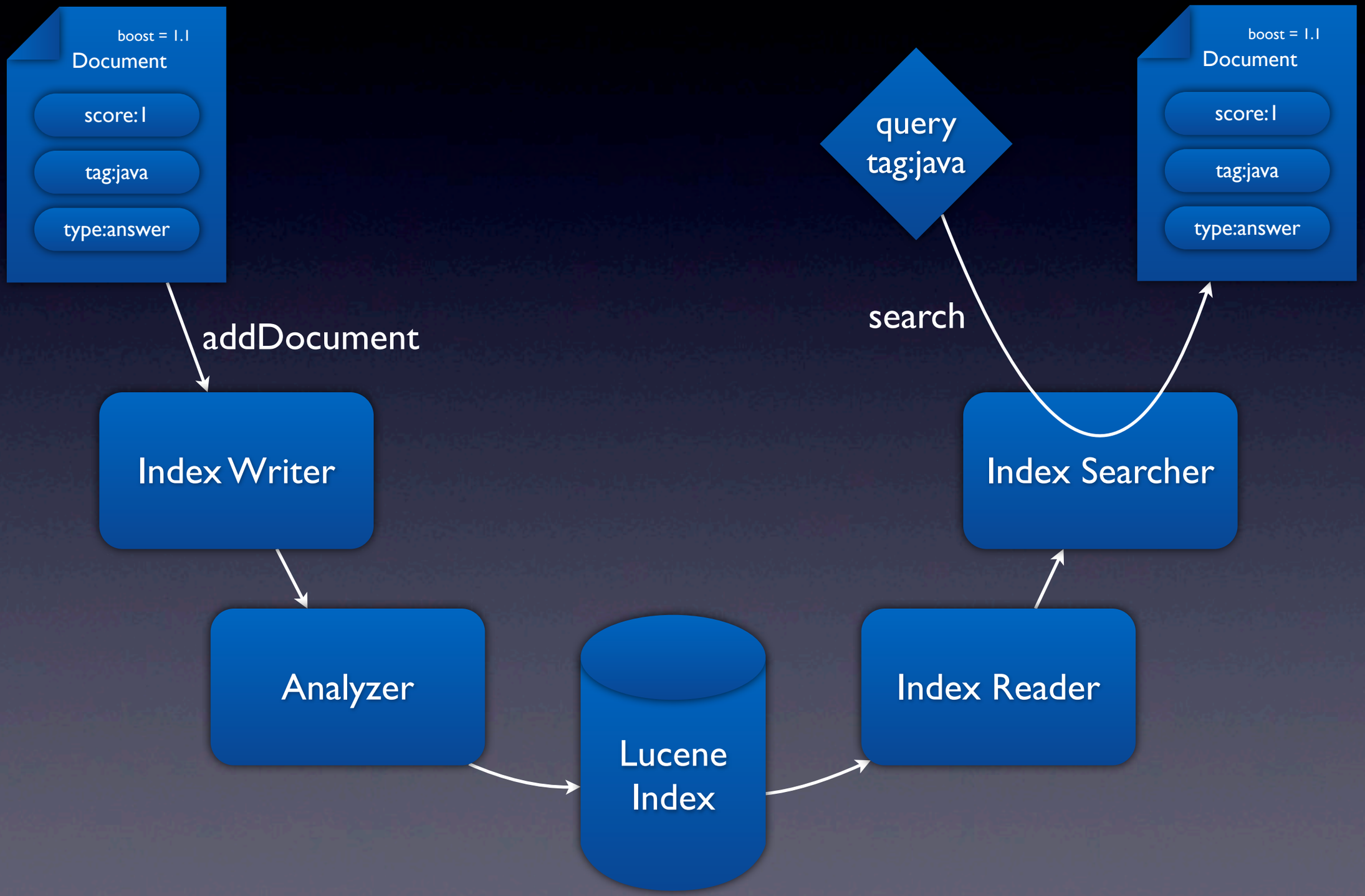
# Basic Flow


boost = 1.1
Document
score:1
tag:java
type:answer
addDocument
Index Writer
Analyzer
Lucene Index
Index Reader
Index Searcher

# Basic Flow

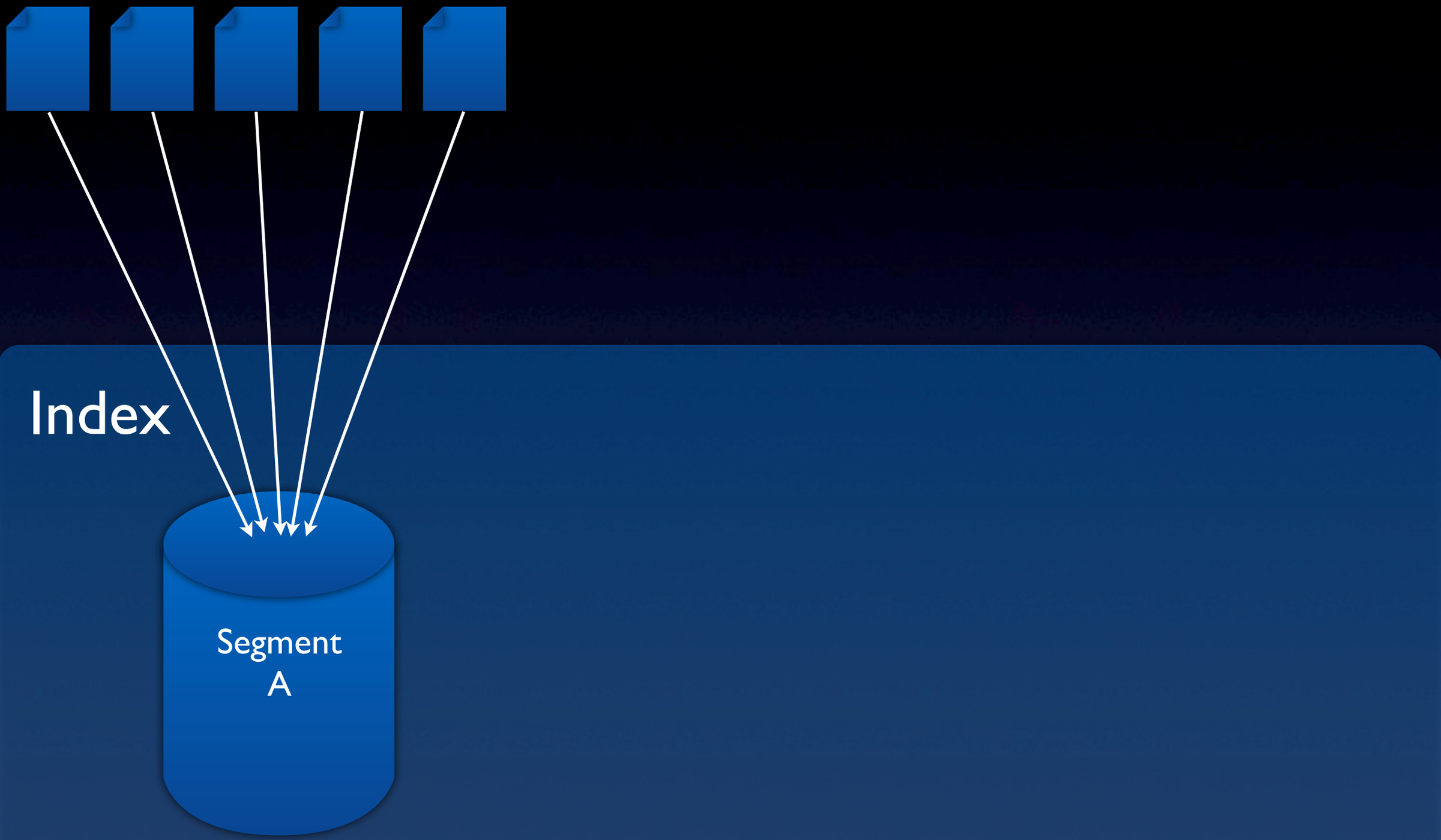

boost = 1.1
Document
score: 1
tag:java
type:answer
addDocument
Index Writer
Analyzer
Lucene Index
Index Reader
Index Searcher
query
tag:java

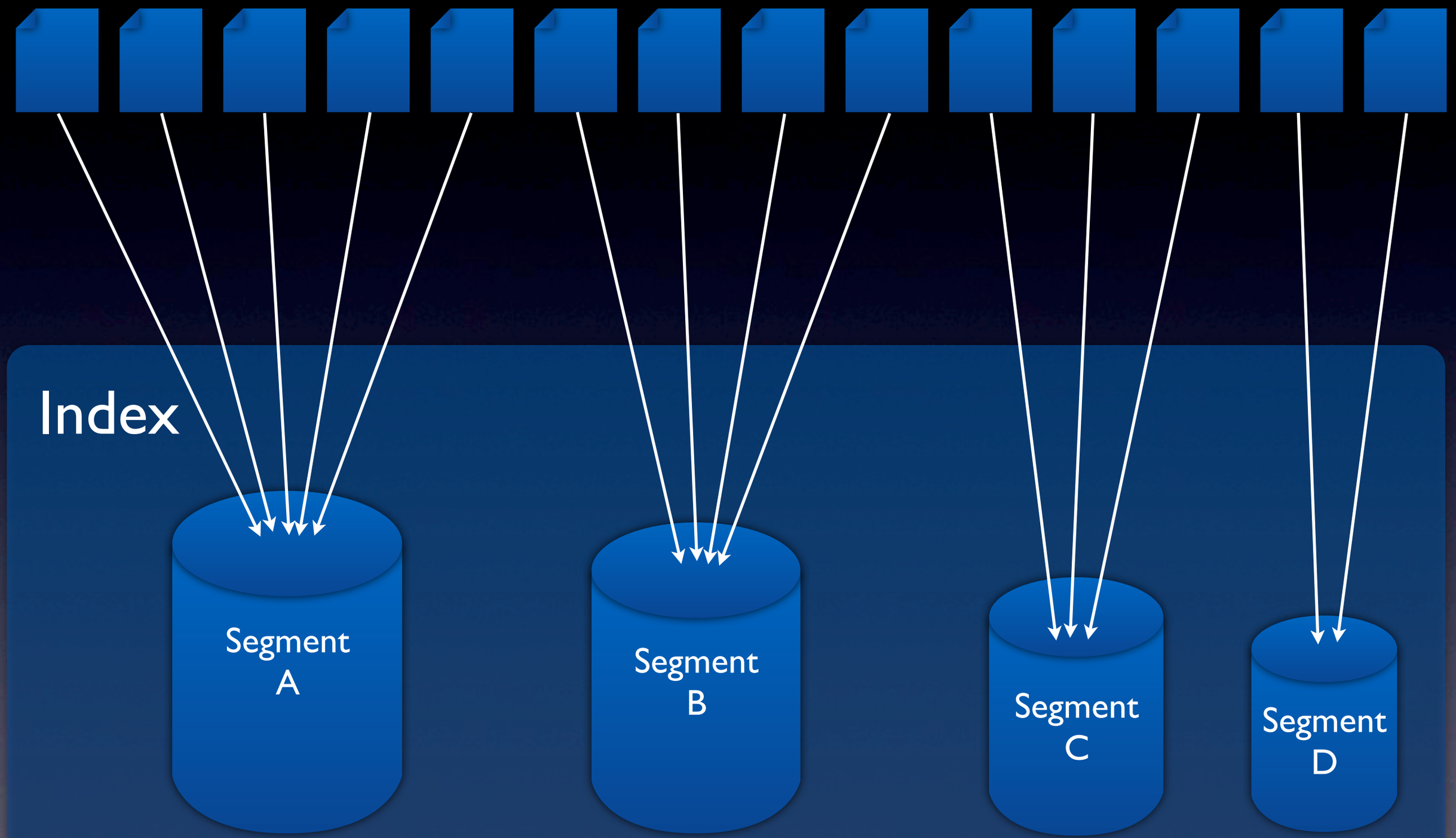
# Basic Flow


boost = 1.1
Document
score:1
tag:java
type:answer
addDocument
Index Writer
Analyzer
Lucene Index
Index Reader
Index Searcher
query tag:java
search
boost = 1.1
Document
score:1
tag:java
type:answer

# Lucene Index Structure

# Index


Index
Segment
A


Index
Segment A
Segment B
Segment C
Segment D

# Term Infos

score:0 3

score:1 4

score:5 2

...

tag:java 2

tag:mysql 3

tag:css 4

...

type:answer 3

type:question 2

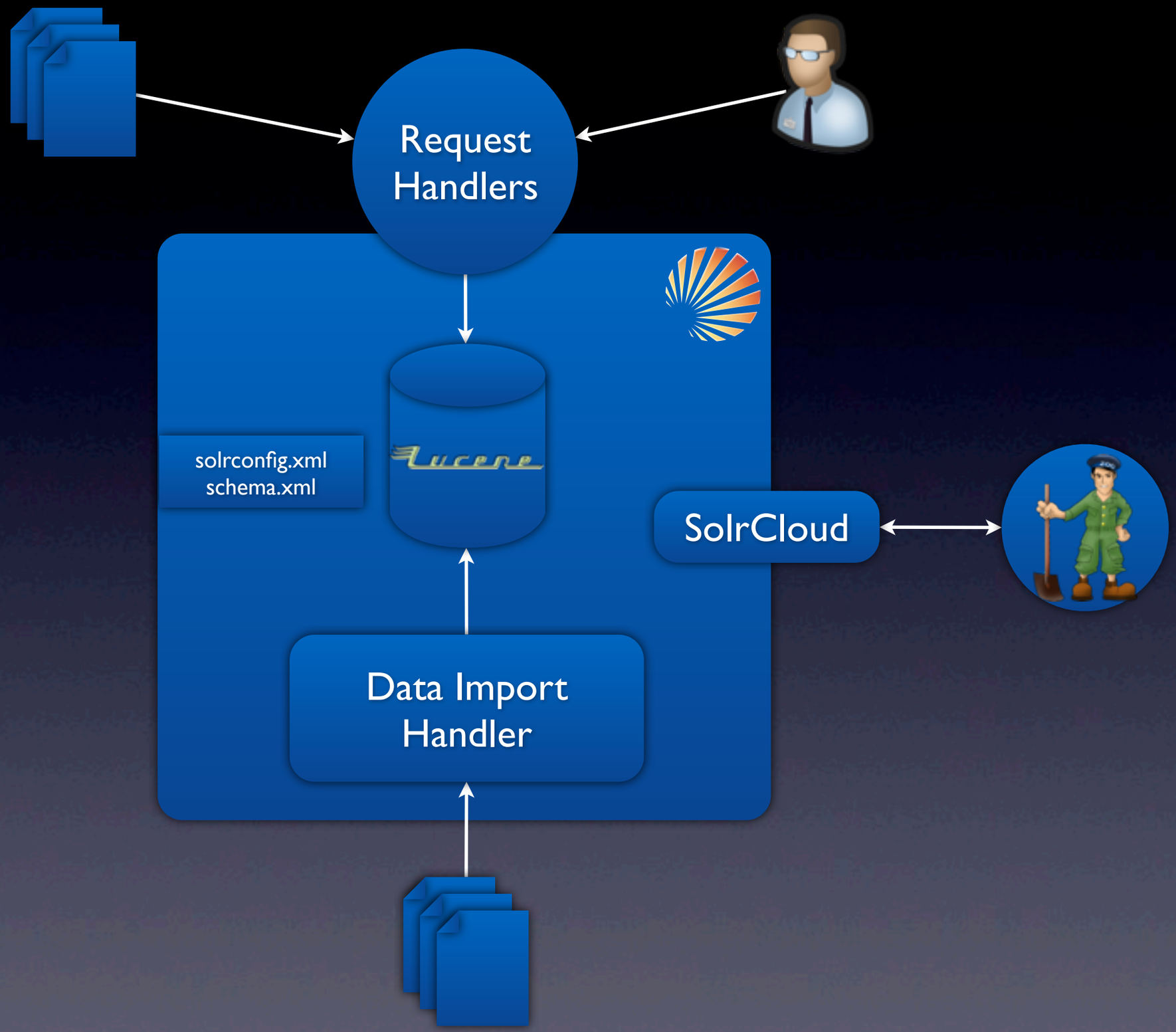
| Term Infos | | Term Frequencies | | | | |
|---|---|---|---|---|---|---|
| score:0 | 3 | 3 | 1 | +1 | 1 | +2 1 |
| score:1 | 4 | 10 | 1 | +3 | 1 | +1 1 +7 1 |
| score:5 | 2 | 4 | 1 | +11 | 1 | |
| ... | | | | | | |
| tag:java | 2 | 5 | 1 | +2 | 1 | |
| tag:mysql | 3 | 6 | 1 | +52 | 1 | +1 1 |
| tag:css | 4 | 1 | 1 | +30 | 1 | +27 1 +2 1 |
| ... | | | | | | |
| type:answer | 3 | 3 | 1 | +7 | 3 | +1 5 |
| type:question | 2 | 5 | 2 | +2 | 1 | |

| Term Info Index | Term Infos | Term Frequencies |
|---|---|---|
| score:0 3 | score:0 3 | 3 1 +1 1 +2 1 |
| | score:1 4 | 10 1 +3 1 +1 1 +7 1 |
| | score:5 2 | 4 1 +11 1 |
| ... | ... | |
| | tag:java 2 | 5 1 +2 1 |
| tag:mysql 3 | tag:mysql 3 | 6 1 +52 1 +1 1 |
| | tag:css 4 | 1 1 +30 1 +27 1 +2 1 |
| | ... | |
| ... | type:answer 3 | 3 1 +7 3 +1 5 |
| type:question 2 | type:question 2 | 5 2 +2 1 |

# Showcase

scalable
enterprise search
server

Apache


Request Handlers
solrconfig.xml
schema.xml
lucene
SolrCloud
Data Import Handler

# SolrCloud

# Join Cluster

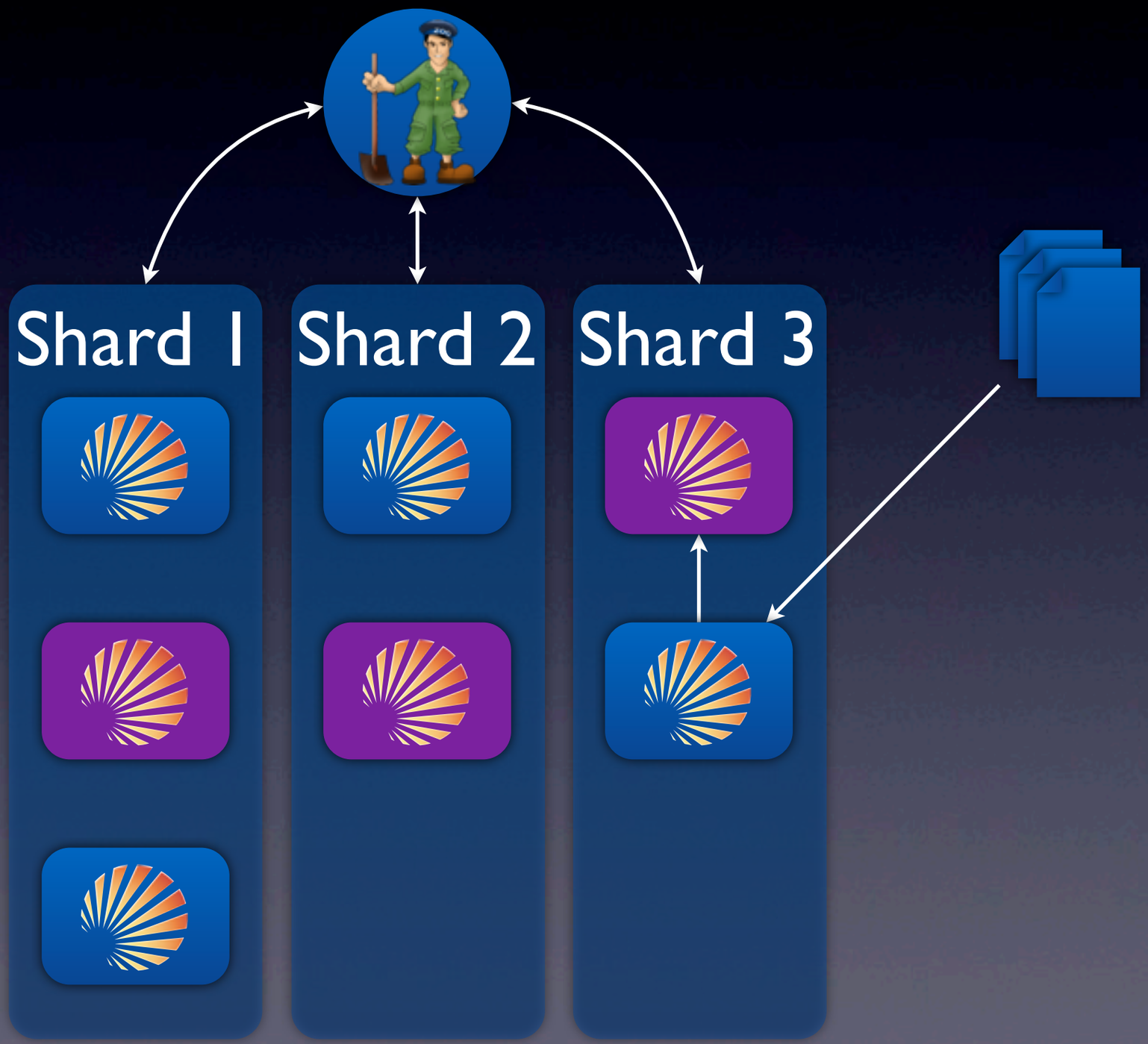
# Join Cluster

# Indexing

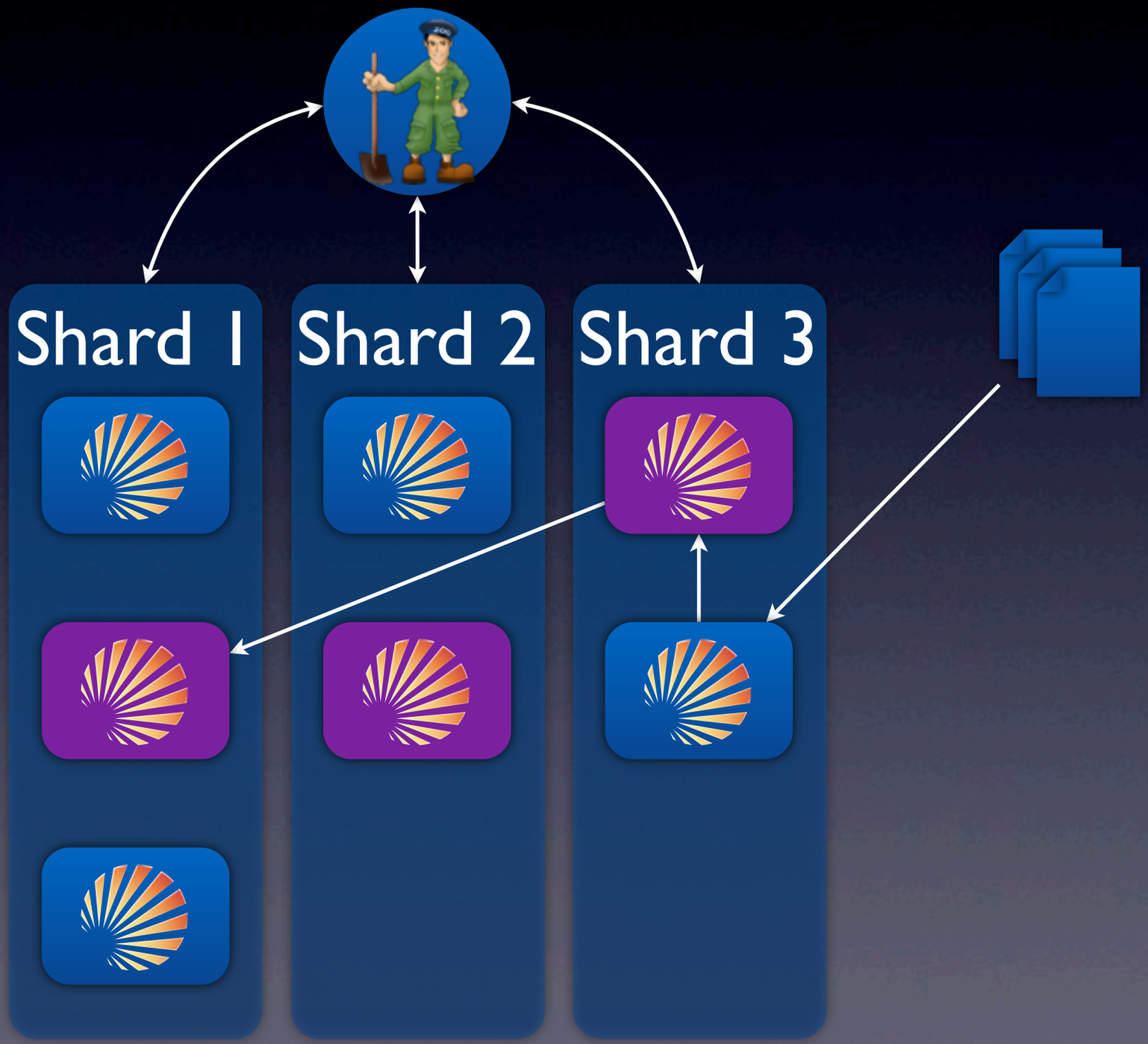
# Indexing

# Indexing

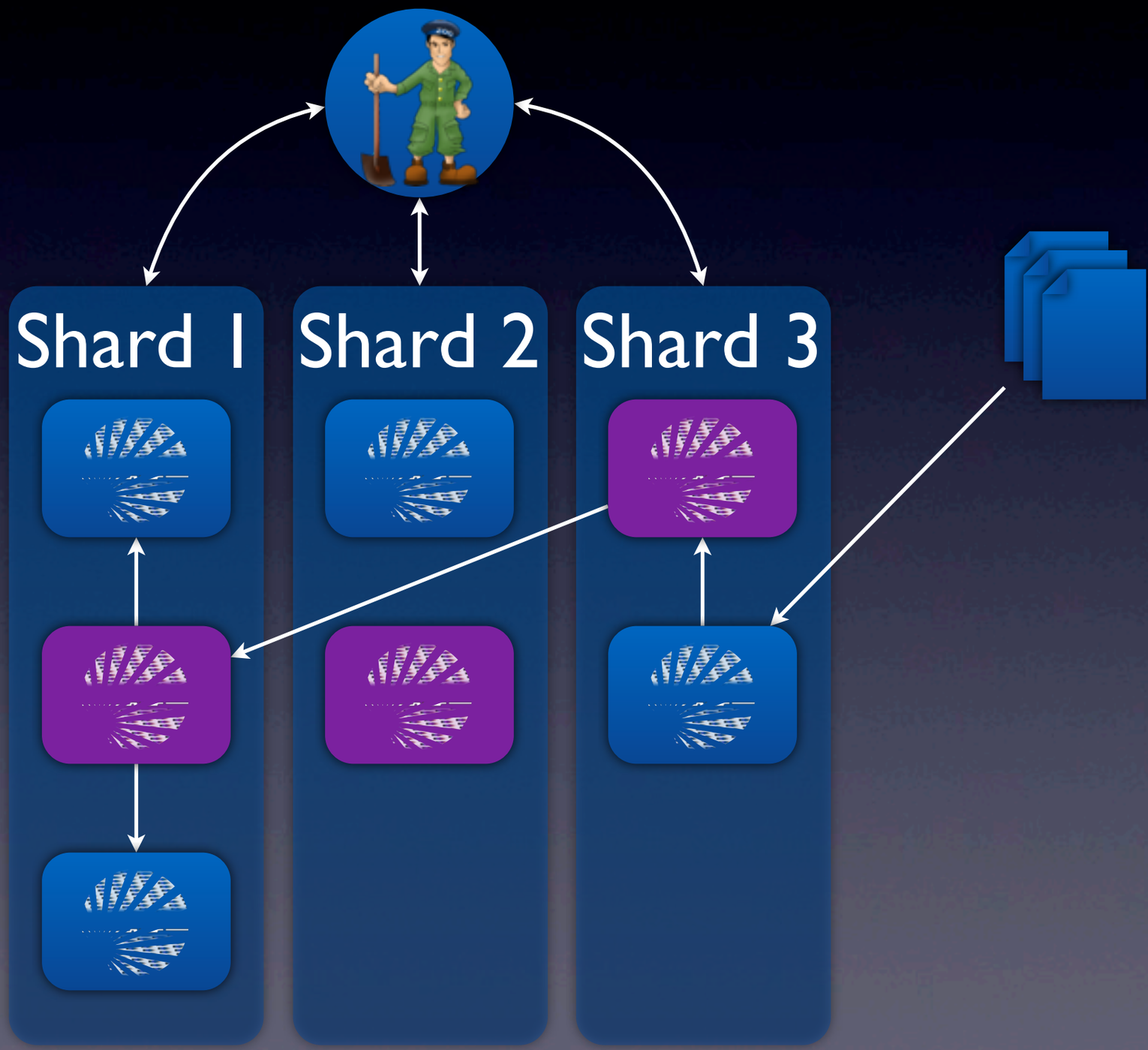
# Indexing



# Indexing



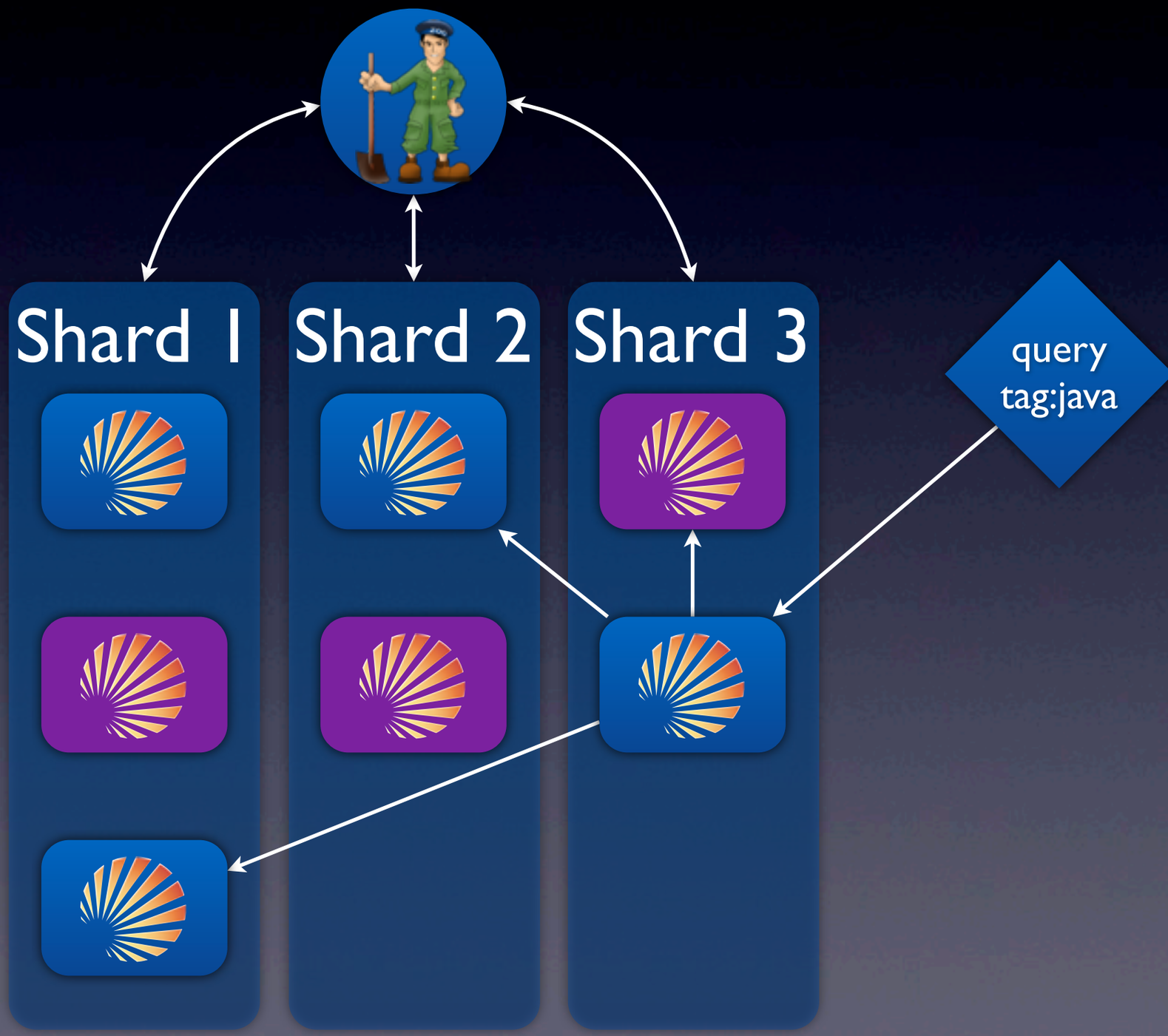
# Indexing



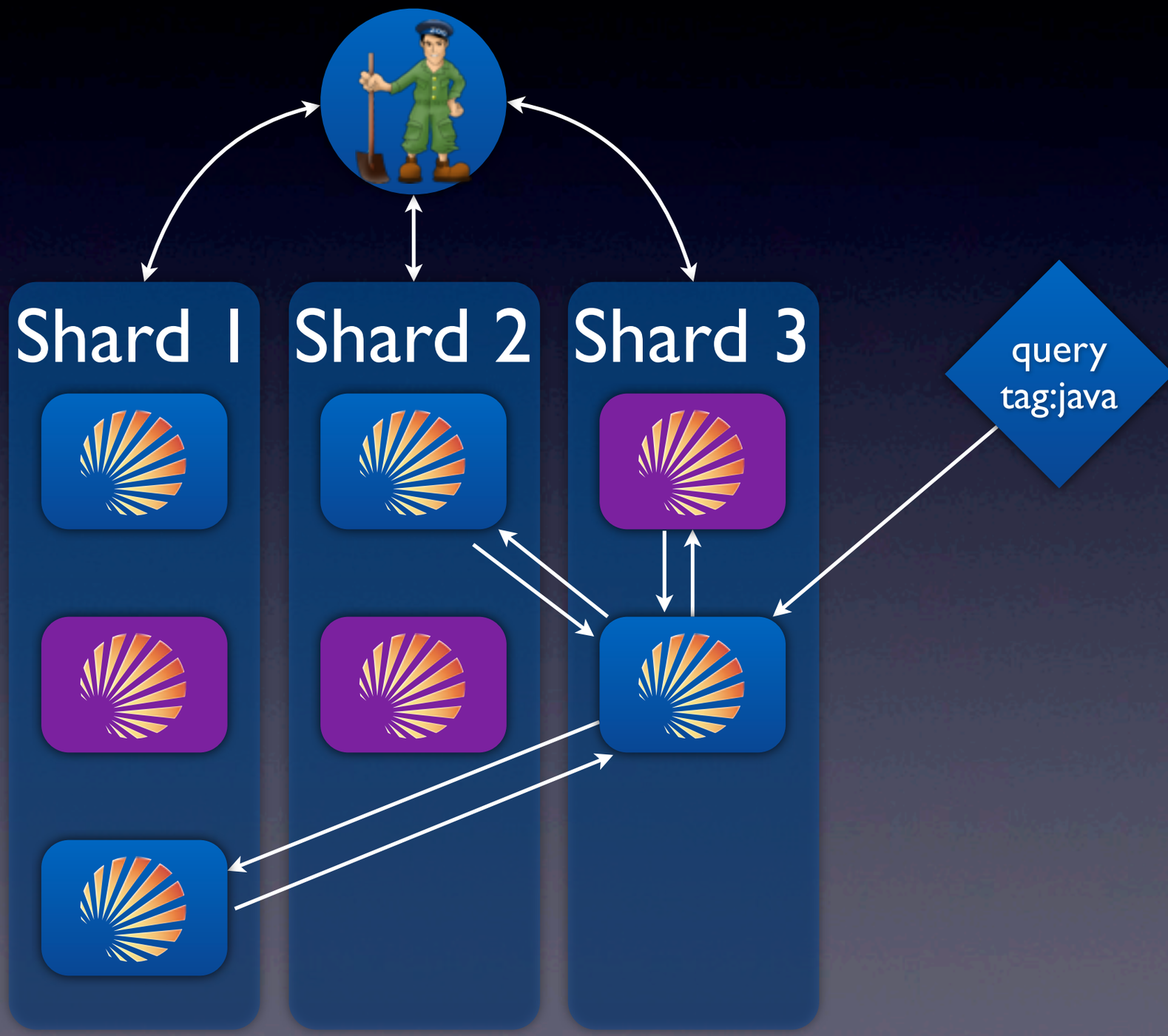
# Query

# Query

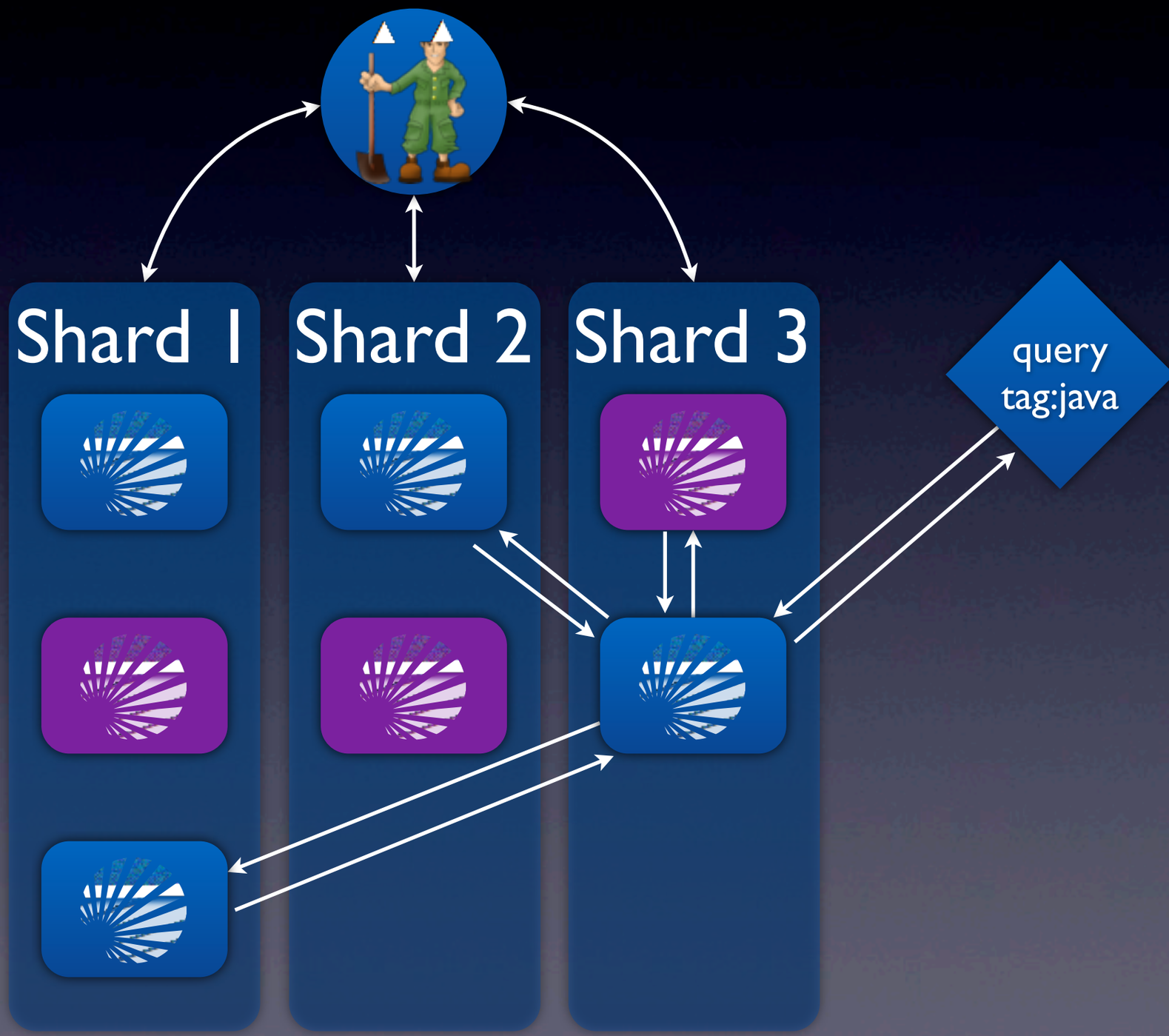
# Query



# Query



# Query

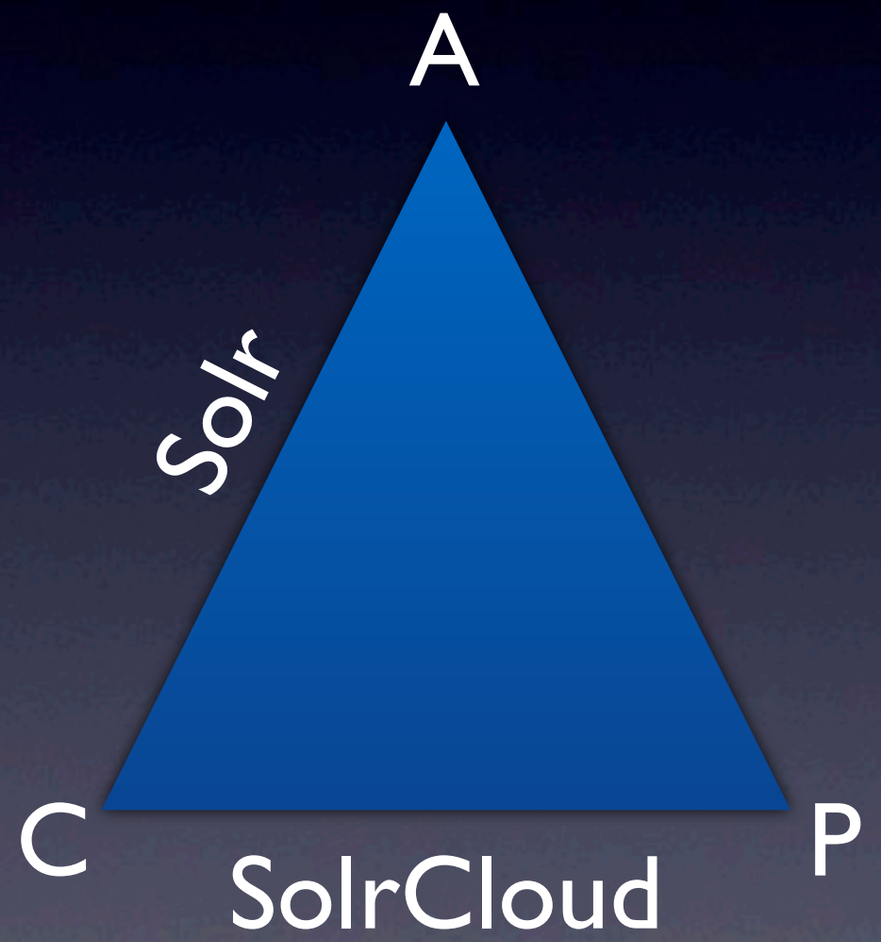


# Failure

# Failure

# Failure

# CAP Model



# Showcase

# Faceted Navigation

# Showcase

# Algorithm

Query Result: 7 31 58 59

Index

| | | | | |
|---|---|---|---|---|
| tag:java | 5 | +2 | | |
| tag:mysql | 6 | +52 | +1 | |
| tag:css | 1 | +30 | +27 | +2 |

# Algorithm

Query Result: 7 31 58 59

Index

| | |
|---|---|
| tag:java | 5 7 |
| tag:mysql | 6 58 59 |
| tag:css | 1 31 58 60 |

# Algorithm

Query Result: 7 31 58 59

Index

| | | | | |
|---|---|---|---|---|
| tag:java | 5 | 7 | | |
| tag:mysql | 6 | 58 | 59 | |
| tag:css | 1 | 31 | 58 | 60 |

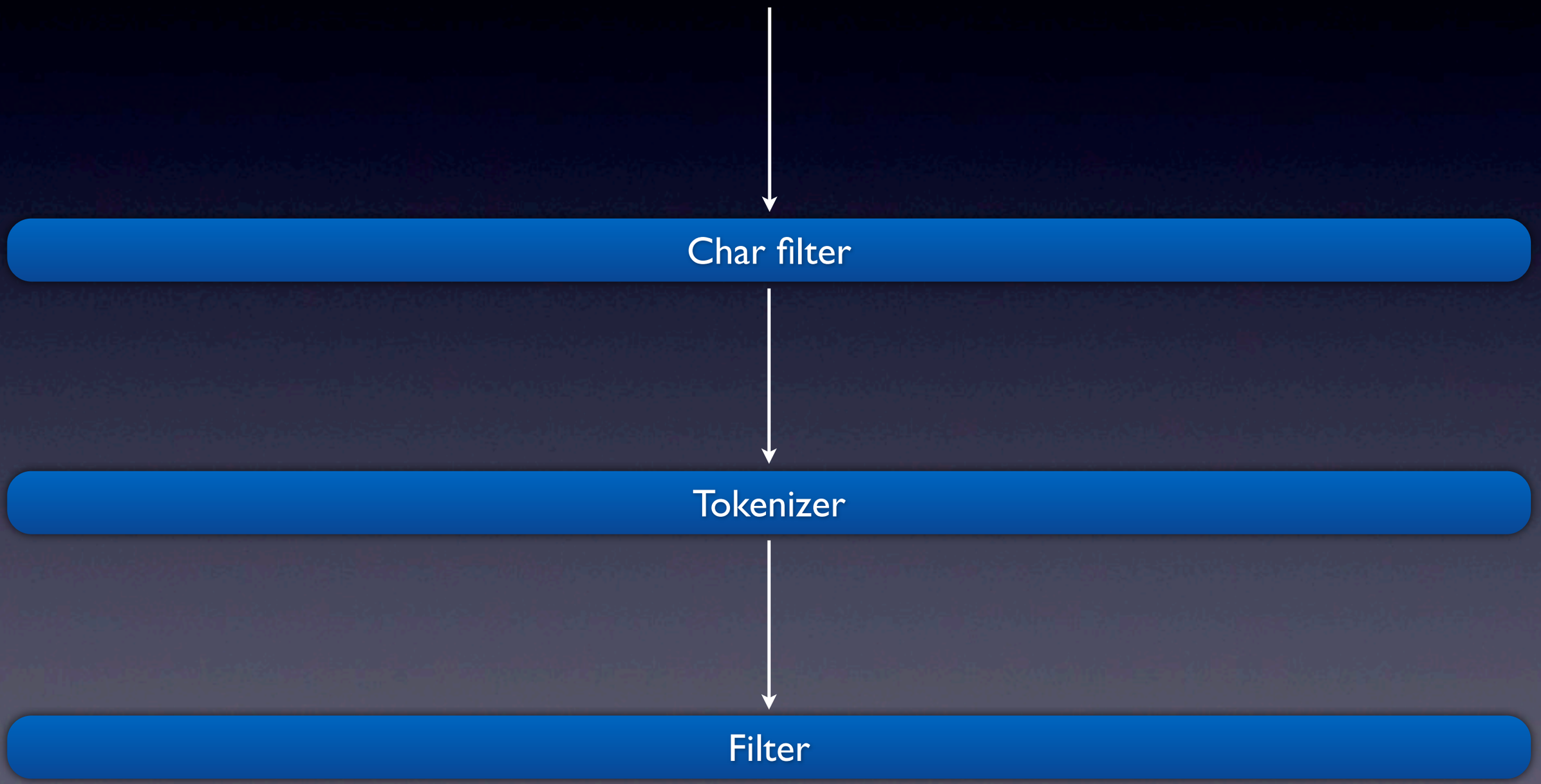
Facet

| |
|---|
| 1 |
| 2 |
| 2 |

# Showcase

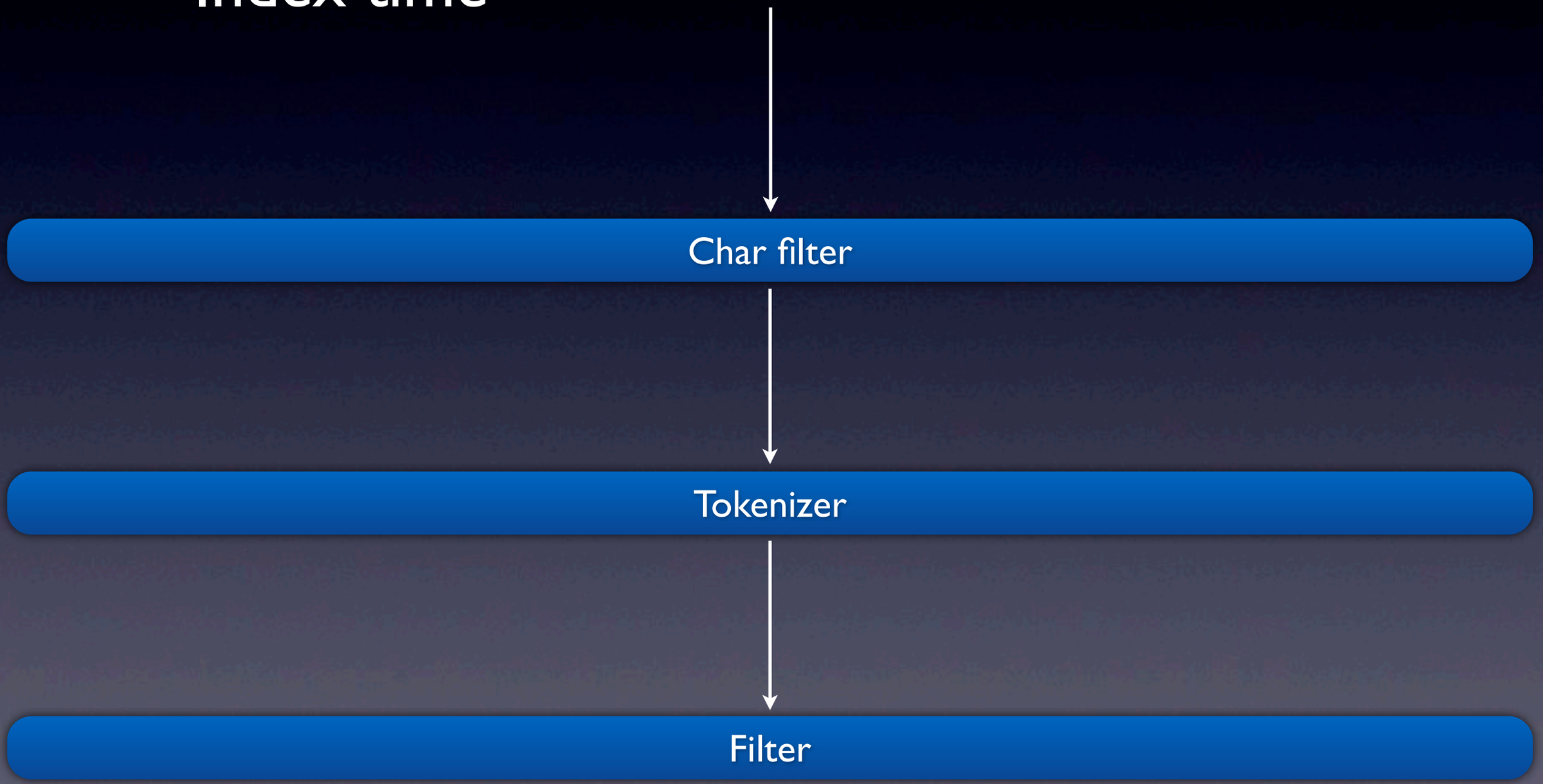
# Text Analysis

# Analyzer



# Analyzer

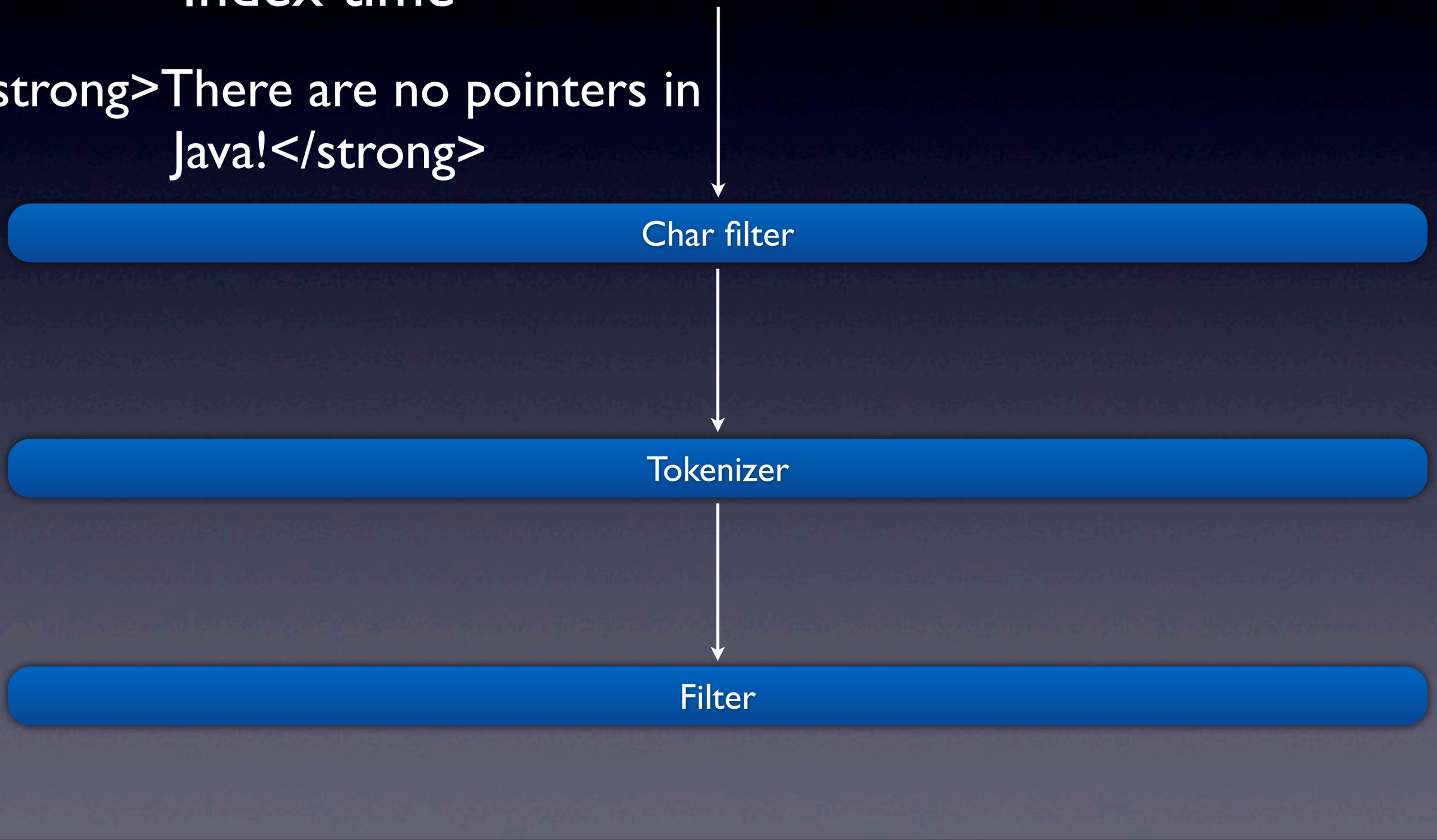
Index time



# Analyzer

Index time

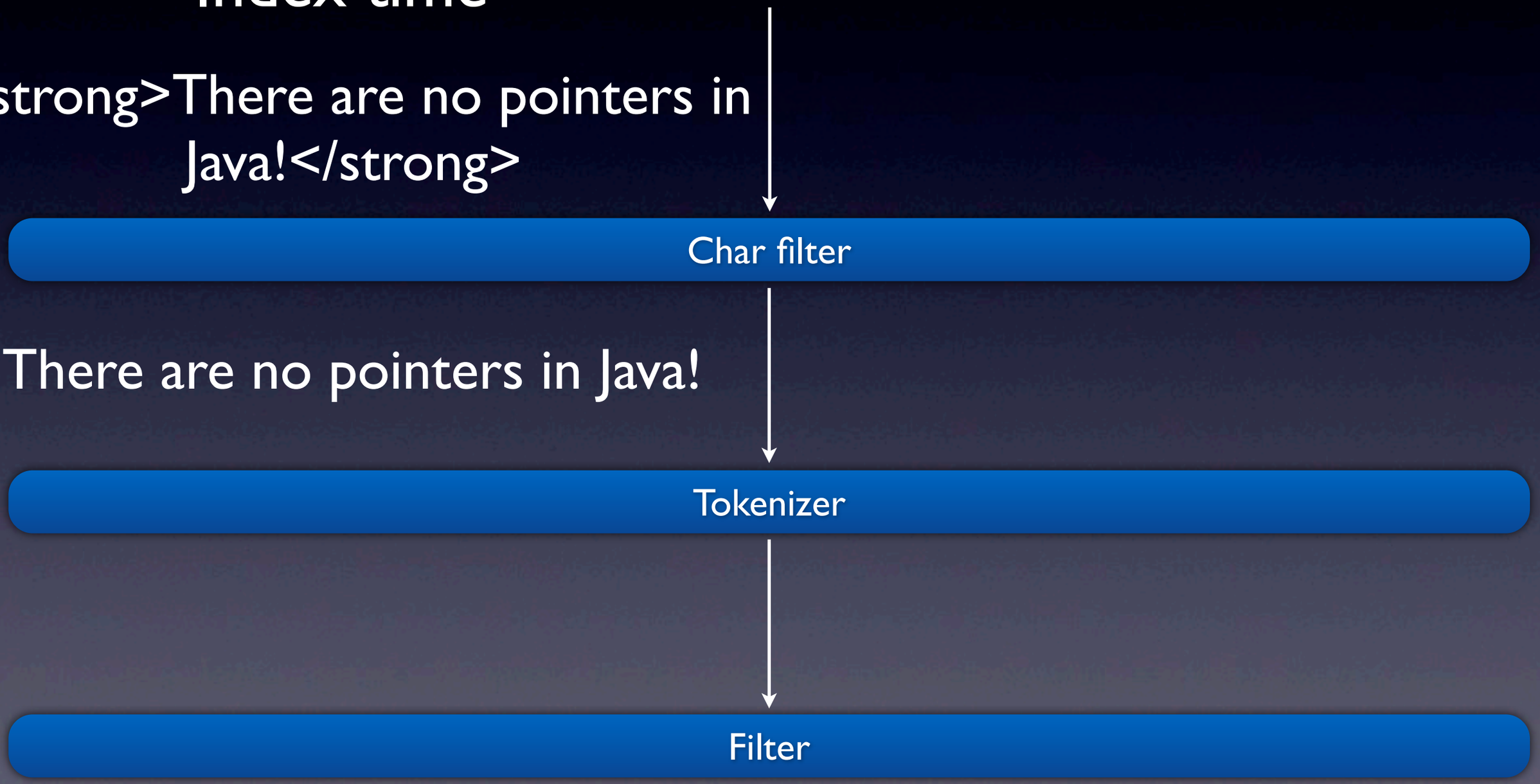
There are no pointers in Java!



# Analyzer

Index time

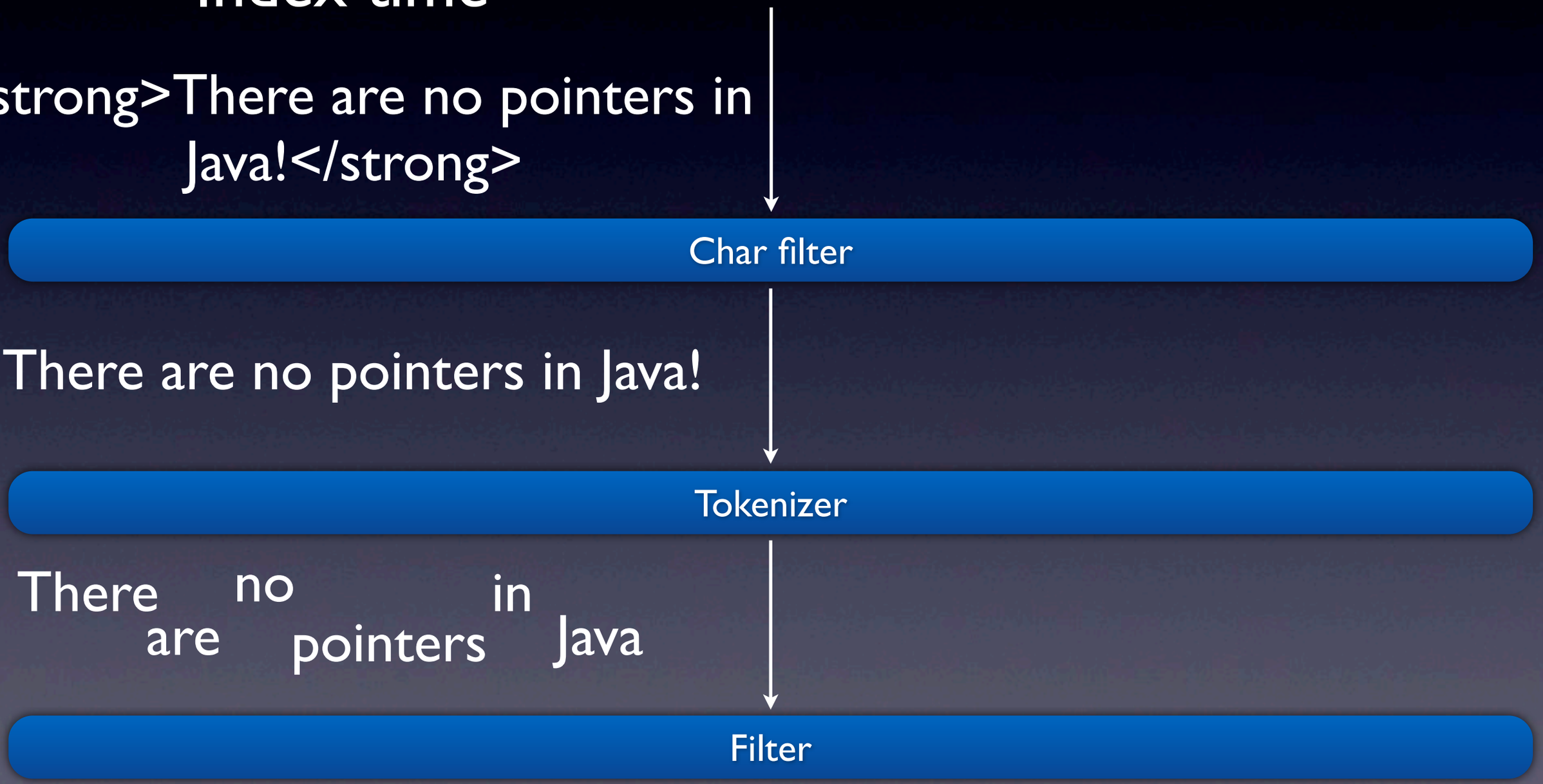
There are no pointers in Java!



# Analyzer

Index time

There are no pointers in Java!



# Analyzer

Index time

There are no pointers in Java!



Char filter

There are no pointers in Java!



Tokenizer

There are no pointers in Java



Filter

? ? ? ? pointer java

# Analyzer

Index time

Query time

There are no pointers in Java!

There are no pointers in Java!

Char filter

Tokenizer

There are no pointers in Java

Filter

? ? ? ? pointer java

# Analyzer

Index time

There are no pointers in Java!

Query time

pointers in Java

Char filter

There are no pointers in Java!

Tokenizer

There are no pointers in Java

Filter

? ? ? ? pointer java

# Analyzer

Index time

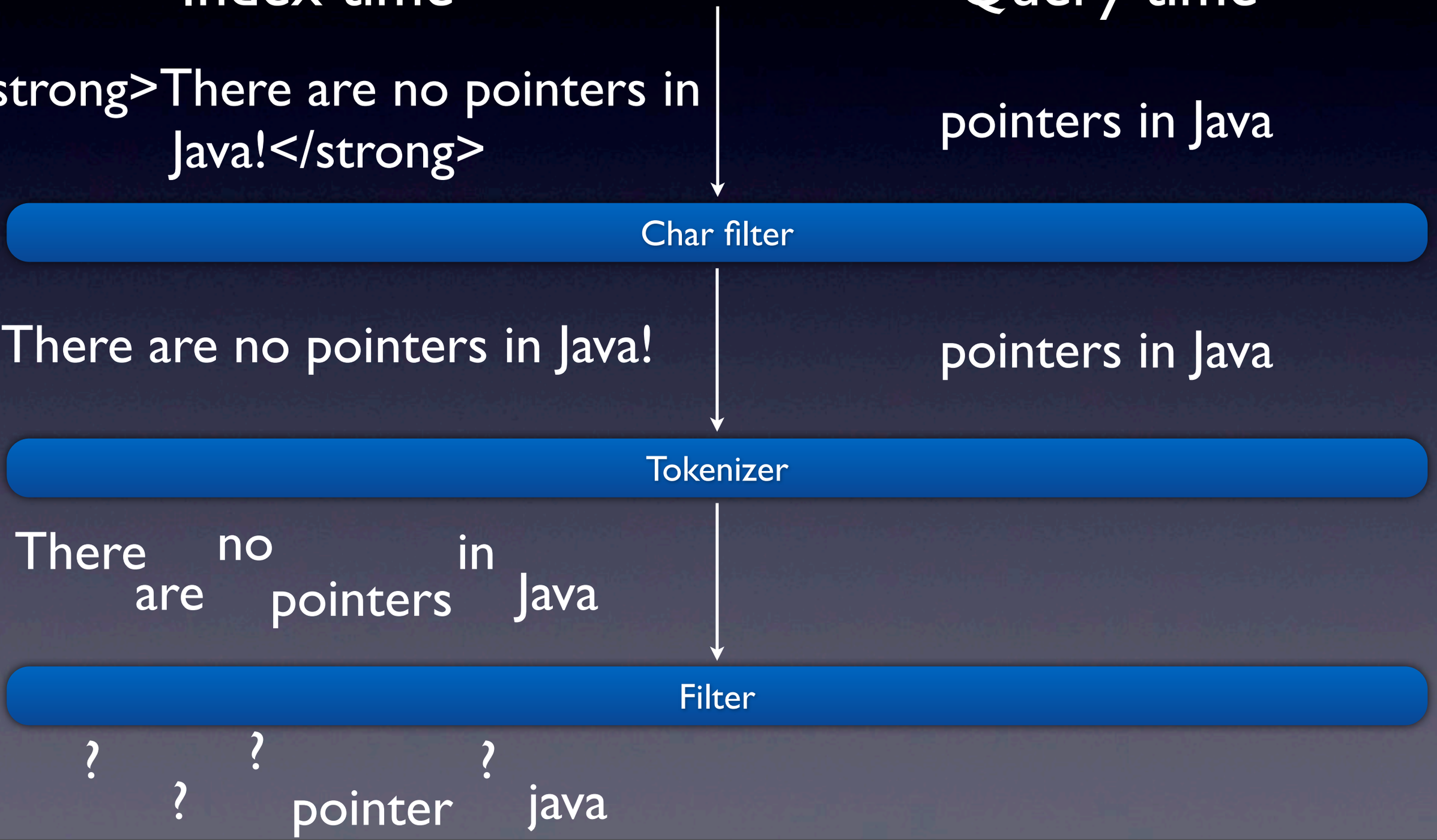
There are no pointers in Java!

There are no pointers in Java!

Query time

pointers in Java

pointers in Java



# Analyzer

Index time

There are no pointers in Java!

Query time

pointers in Java

Char filter

There are no pointers in Java!

pointers in Java

Tokenizer

There are no pointers in Java

pointers in Java

Filter

? ? ? ? pointer java

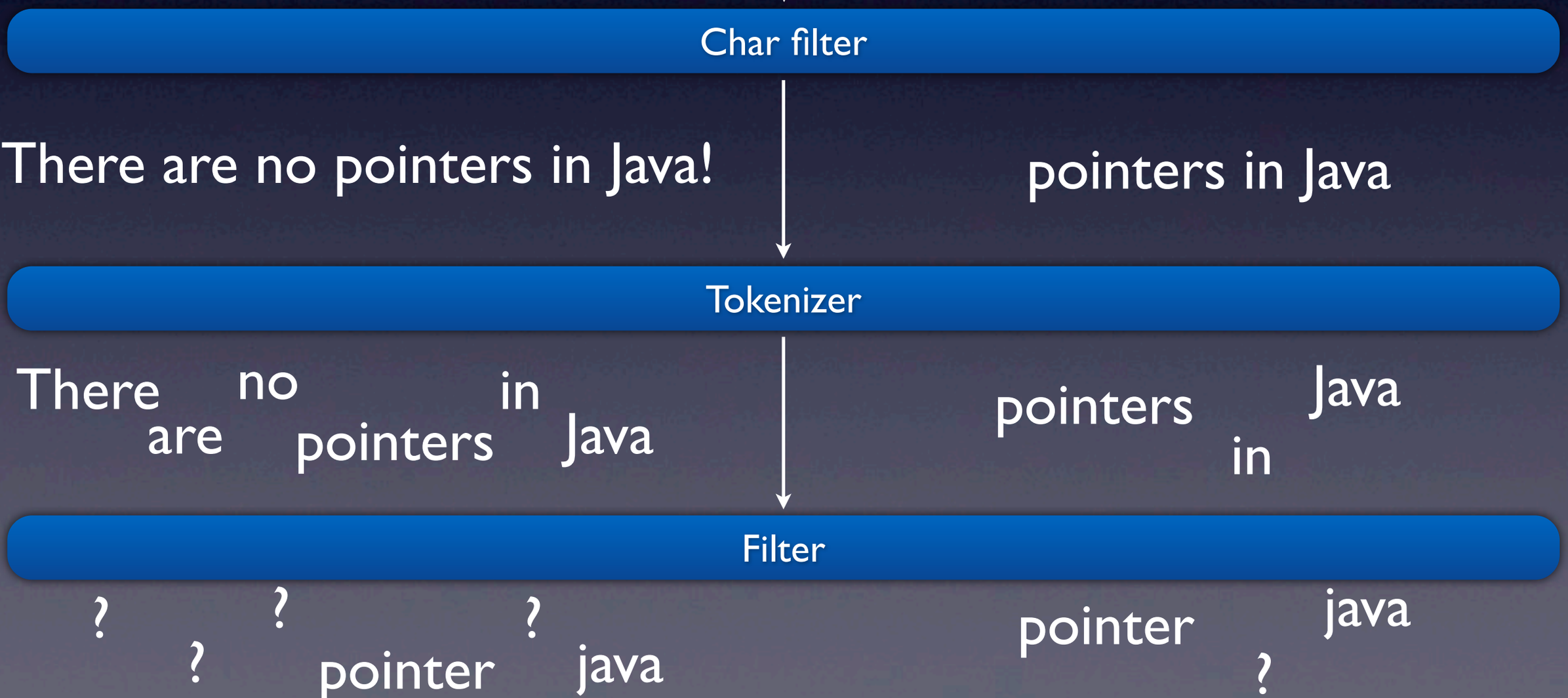
# Analyzer

Index time

There are no pointers in Java!

Query time

pointers in Java



# Showcase

# Spell Suggestions

# Levenshtein Distance

Levenshtein distance = 1

html

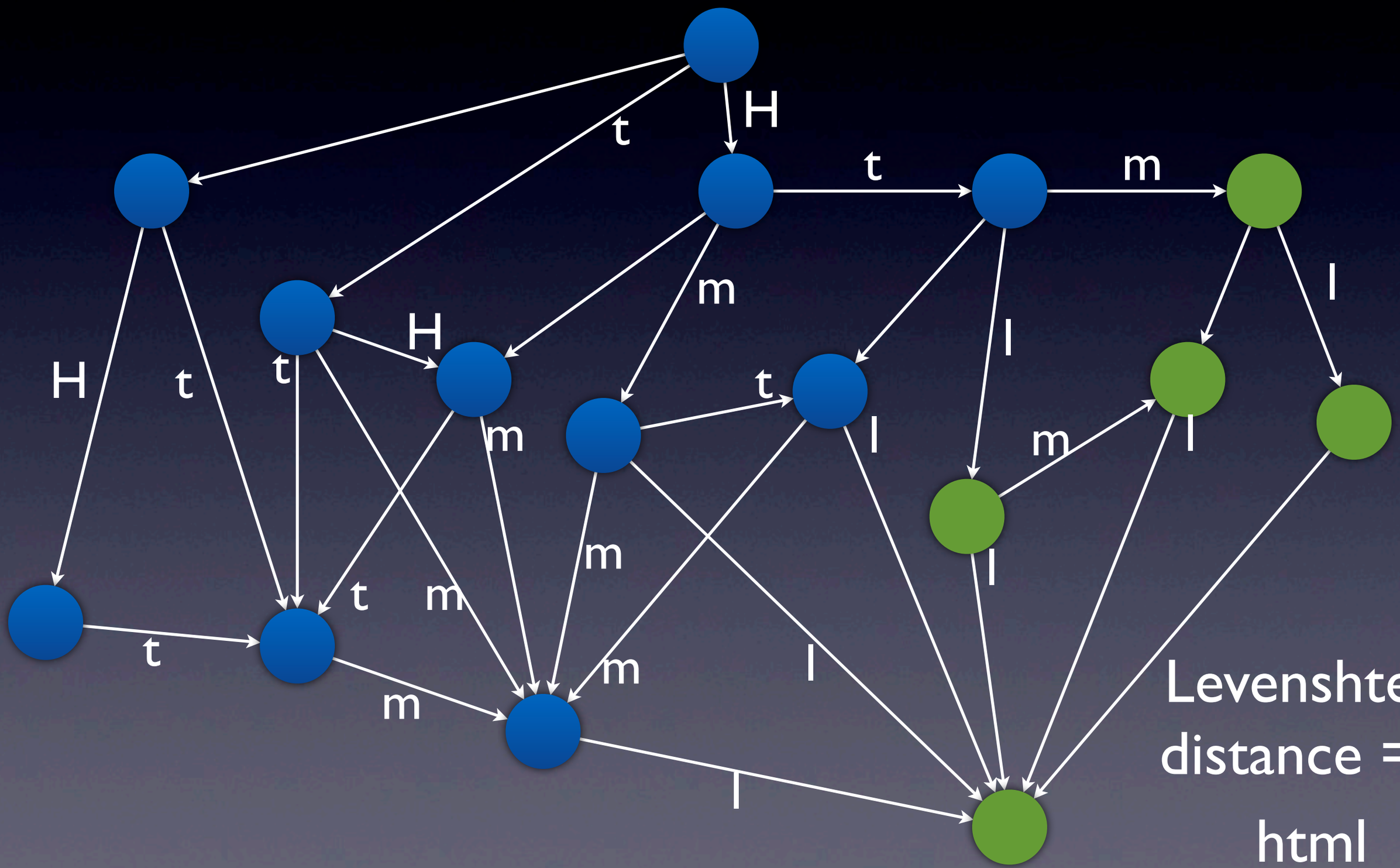
htmm

Levenshtein distance = 2

html

hlmz

# Levenshtein Automaton



# Showcase

# Solr is...

- enterprise level search engine
- vertically scalable
- horizontaly scalable, but...
- tunable
- poorly documentation
- with active community

# References

- http://blog.mikemccandless.com
- http://lucene.apache.org/core/4_3_1/index.html

- Introduction to Information Retrieval http://nlp.stanford.edu/IR-book/
- http://wiki.apache.org/solr/
- https://cwiki.apache.org/confluence/display/solr/Apache+Solr+Reference+Guide

# Q&A


http://flip.it/whFqy

asokolenko@griddynamics.com

http://twitter.com/AnatolSokolenko

# The End