



# ТЕХНОСФЕРА

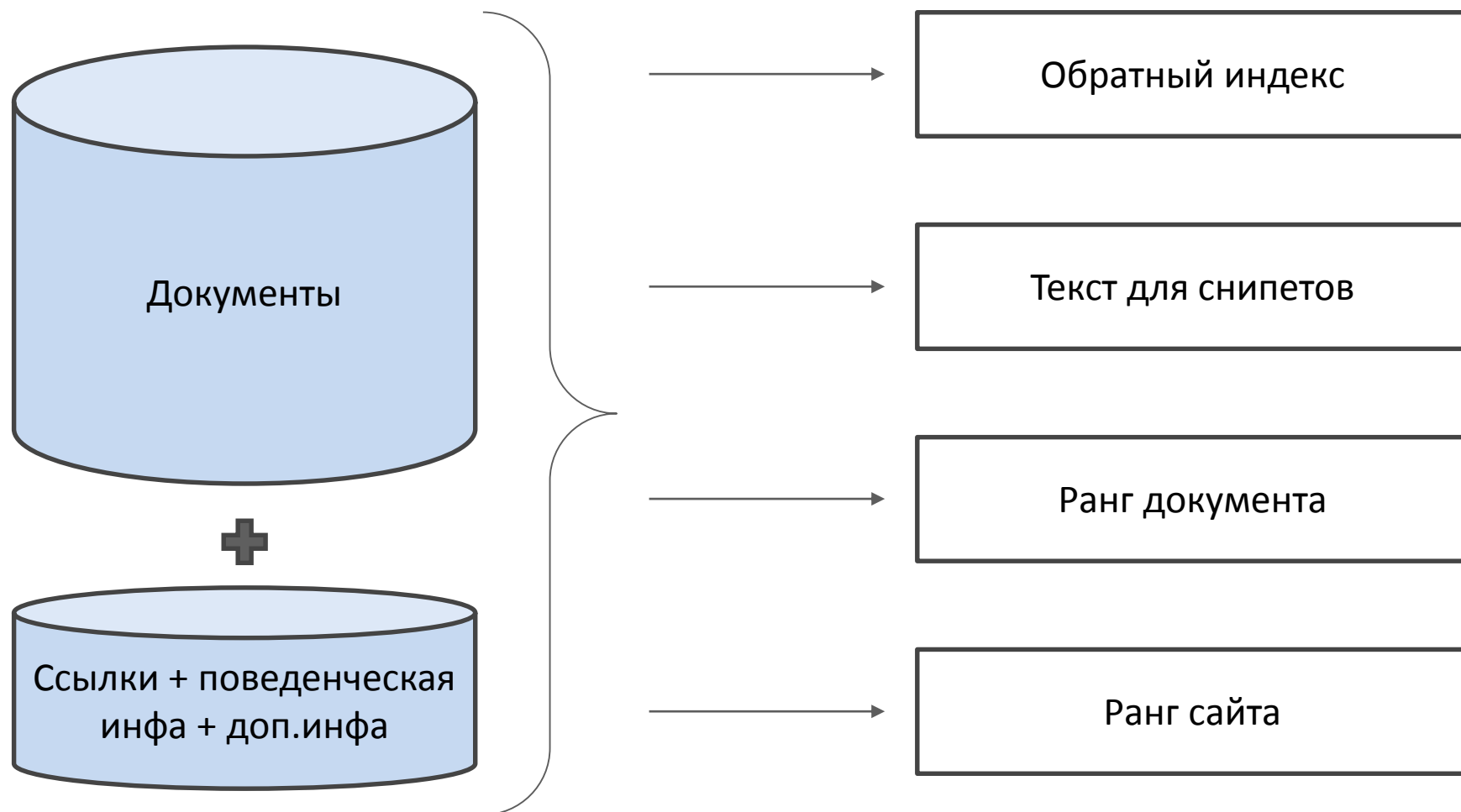
## Индексация

Сергукова Юлия,  
программист отдела инфраструктуры проекта  
Поиск@Mail.Ru

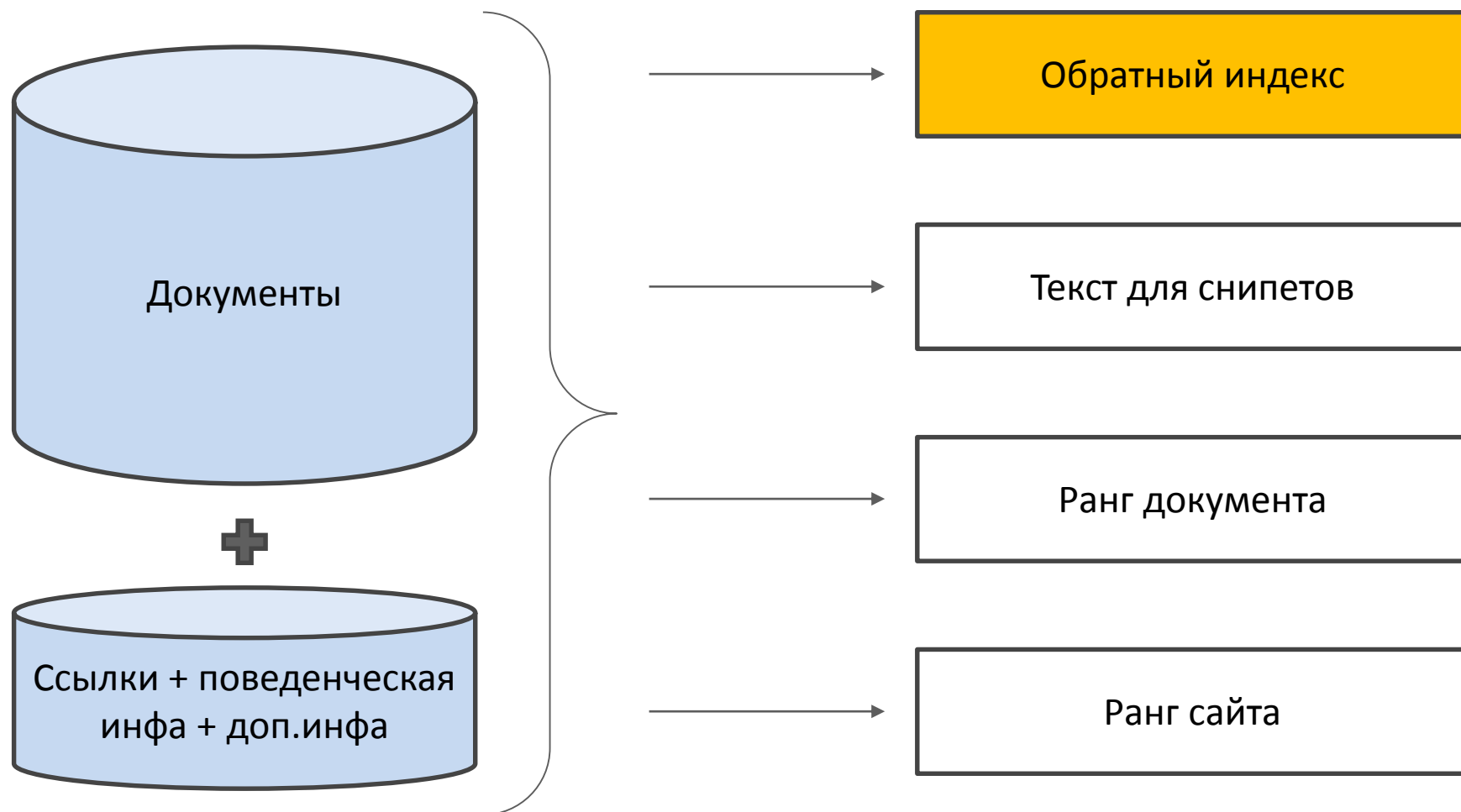
## План лекции:

- 1. Обратный индекс**
- 2. Поиск по обратному индексу, пересечение блоков**

# Что мы получаем из документов



# Что мы получаем из документов



# От документа к индексу

1. Документ  $\rightarrow$  текст

# От документа к индексу

1. Документ → текст
2. Текст → слова

# От документа к индексу

1. Документ → текст
2. Текст → слова
3. Слова → леммы (зачем?)

# От документа к индексу

1. Документ → текст
2. Текст → слова
3. Слова → леммы (зачем?)

Лемма – лингв. нормализованная, основная форма слова, вместе с информацией о построении других форм



# От документа к индексу

1. Документ → текст
2. Текст → слова
3. Слова → леммы (зачем?)
4. Лемма → список документов, в которых встречается:  
«posting list», «инвертированный список» и т.д.

# От слов к ~~действию~~ числам

Документ <-> URL

# От слов к ~~действию~~ числам

Документ  $\leftrightarrow$  URL  $\leftrightarrow$  docID

# От слов к ~~действию~~ числам

Документ  $\leftrightarrow$  URL  $\leftrightarrow$  docID

Слово  $\leftrightarrow$  termID (например, hash)

# Обратный индекс



# Чем дополнить обратный индекс?



# Чем дополнить обратный индекс?

1. Ранг термина
2. Координаты

Нужные данные для поиска и ранжирования  
Мы ограничены размером и скоростью

# Быстрый и компактный

## 1. Быстрый:

1. Больше нагрузка – все запросы
2. Пользователь не будет ждать!

## 2. Компактный:

1. Завязано на скорость – можем хранить в RAM

+

## Гибкий:

- Хранить разные данные (зонные индексы)
- Масштабируемый / разделяемый



## Физические ограничения

1. RAM быстрее HDD на 2-3 порядка
2. SSD? Быстро, но пока недостаточно надежно для ВНС
3. Гибриды – дорого и малый объем
4. RAM ограничена по объему

Скорость меряется в IOpS – Input/Output per Second

# Память: как правильно с ней работать?

## 1. Считать блок:

- Спозиционироваться
- Считать

## 2. Что быстрее?

1. Считать 1GB, записанный непрерывно
2. Считать 1024 блока по 1MB
3. Считать  $1024 * 1024$  блока по 1KB

# Память: как правильно с ней работать?

1. Меньше позиционируемся – больше читаем
2. Меньший объем данных – меньше читать

# Наша конфигурация

Компонент	тестовый кластер
CPU	Xeon: 2x8 core, HT*; 2.4 Ghz
RAM	48 Gb
Диски	1Tb+ SATA
Скорость	~10ms

\* Hyper Threading – технология, «честно» реализующая параллельную работу 2 тредов

# Размер индекса

1. 10кМ документов
2. 1 документ  $\sim$  70Kb
3. 1 лемма  $\sim$  8b

Проблема:

очень большой словарь

# Разделяемый индекс

Как поделить большой индекс между несколькими серверами?



# Разделяемый индекс

1. Сервер  $\leftrightarrow$  терм
2. Сервер  $\leftrightarrow$  документ

# Бинарный поиск

$A \& B$

$A || B$

$\neg B$



# Пересечение блоков





pathfinder kingmaker


×


Найти


 Везде

 Картинки

 Видео

 Приложения

 Новости

 Ответы

# Пересечение блоков

 × Найти[Везде](#)[Картинки](#)[Видео](#)[Приложения](#)[Новости](#)[Ответы](#)

pathfinder

94

7

12

55

57

43

kingmaker

94

1

7

# Пересечение блоков

 × Найти[Везде](#)[Картинки](#)[Видео](#)[Приложения](#)[Новости](#)[Ответы](#)

pathfinder

94

7

12

55

57

43

kingmaker

94

1

7

# Пересечение блоков

 × Найти[Везде](#)[Картинки](#)[Видео](#)[Приложения](#)[Новости](#)[Ответы](#)

pathfinder

7

12

43

55

57

94

kingmaker

1

7

94

Пройти по всему списку – долго

# Пересечение блоков

 × Найти

Везде

Картинки

Видео

Приложения

Новости

Ответы

pathfinder



kingmaker

Jump Tables!

# Пересечение блоков

 × Найти

Везде

Картинки

Видео

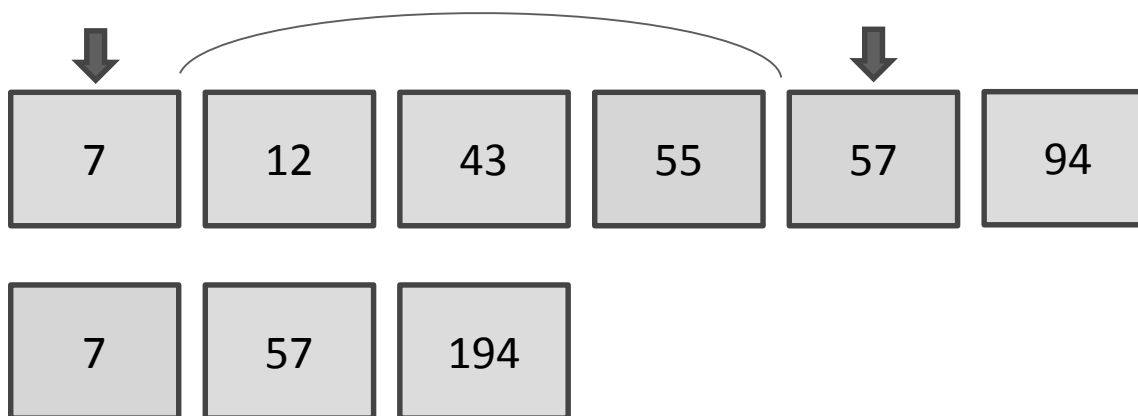
Приложения

Новости

Ответы

pathfinder

IT



Jump Tables!

# Пересечение блоков



x

Найти

[Везде](#)[Картинки](#)[Видео](#)[Приложения](#)[Новости](#)[Ответы](#)

pathfinder

7

12

43

55

57

94

JT

7

57

194

JT2

7

243

