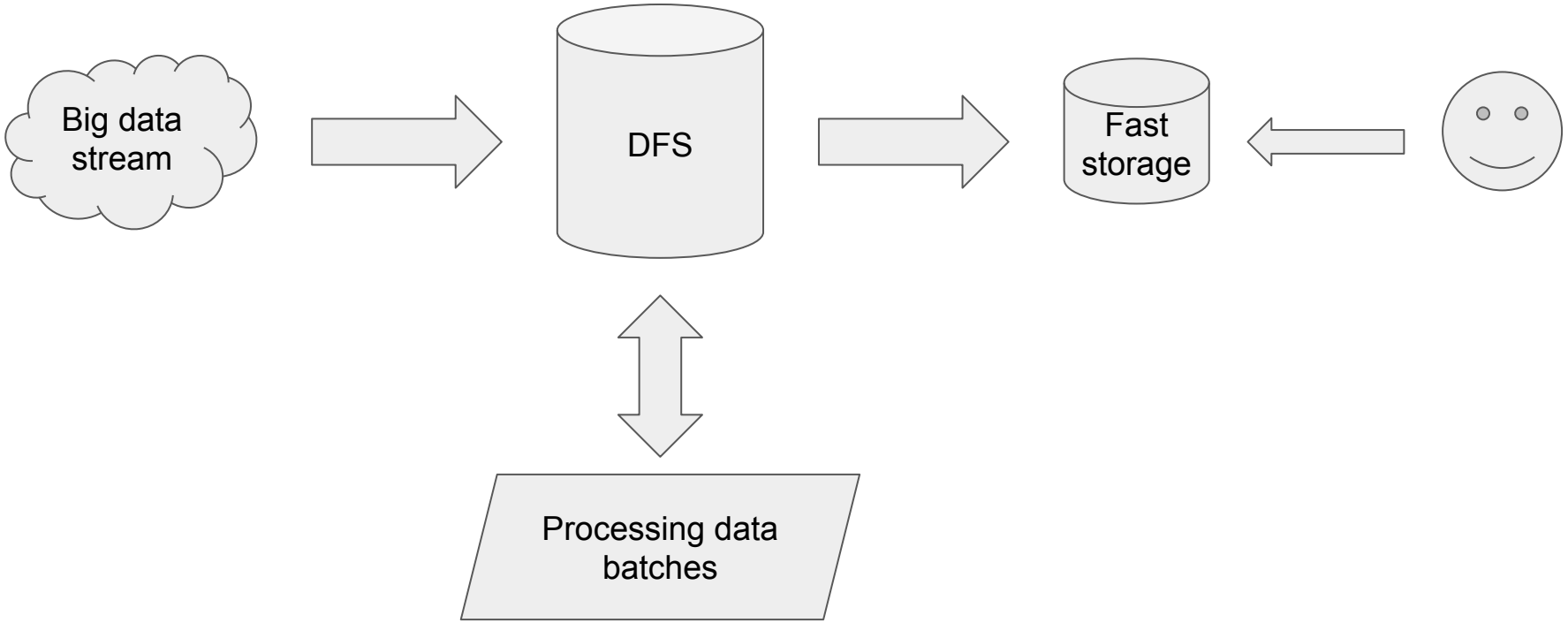
# Realtime Big Data Intro

**Vybornov Artyom**, avybornov@bigdatateam.org
Big Data Instructor, http://bigdatateam.org/
Head of Data Platform, Rambler Group
https://www.linkedin.com/in/artvybor/

BIGDATA TEAM

# Батчевый подход



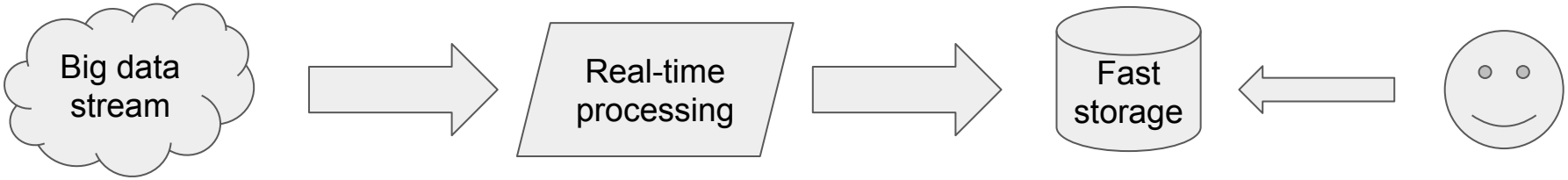
BIGDATA TEAM

# Главный минус батчевого подхода

- Лаг (задержка) это время которое между возникновением события и моментом, когда оно повлияло на результат решения задачи
  - На практике батч это большой интервал времени (обычно день или час)
  - Размер батча определяет минимальное значение лага
- Для многих задач работает правило: чем меньше лаг, тем более ценные данные мы получили

# Real-time big data

- Real-time big data - набор технологий обработки Big Data с минимально возможным лагом
- Ключевые особенности:
  - Без DFS
  - Работа не с батчем, а с потоком событий



# Real-time бывает разным

- Лаг в минуты
  - Ранжирование ленты новостей под пользователя (Facebook, Vkontakte, Yandex.Dzen)

# Real-time бывает разным

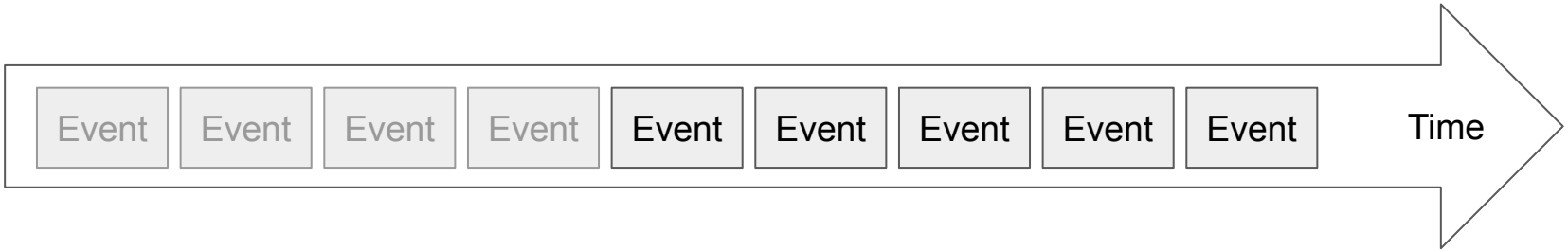
- Лаг в минуты
  - Ранжирование ленты новостей под пользователя (Facebook, Vkontakte, Yandex.Dzen)
- Лаг в секунды
  - Современная RTB Реклама (Google, Yandex, Rambler)

# Real-time бывает разным

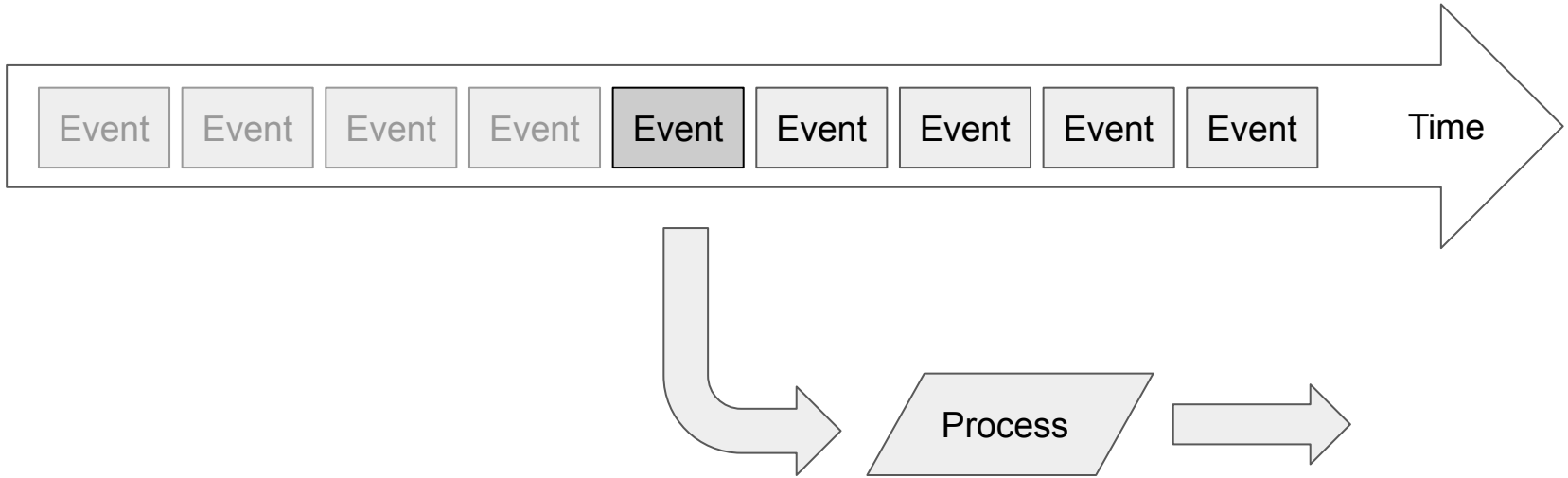
- Лаг в минуты
  - Ранжирование ленты новостей под пользователя (Facebook, Vkontakte, Yandex.Dzen)
- Лаг в секунды
  - Современная RTB Реклама (Google, Yandex, Rambler)
- Лаг в миллисекунды
  - High-frequency trading (HFT)

BIGDATA TEAM

# Пособытийный подход

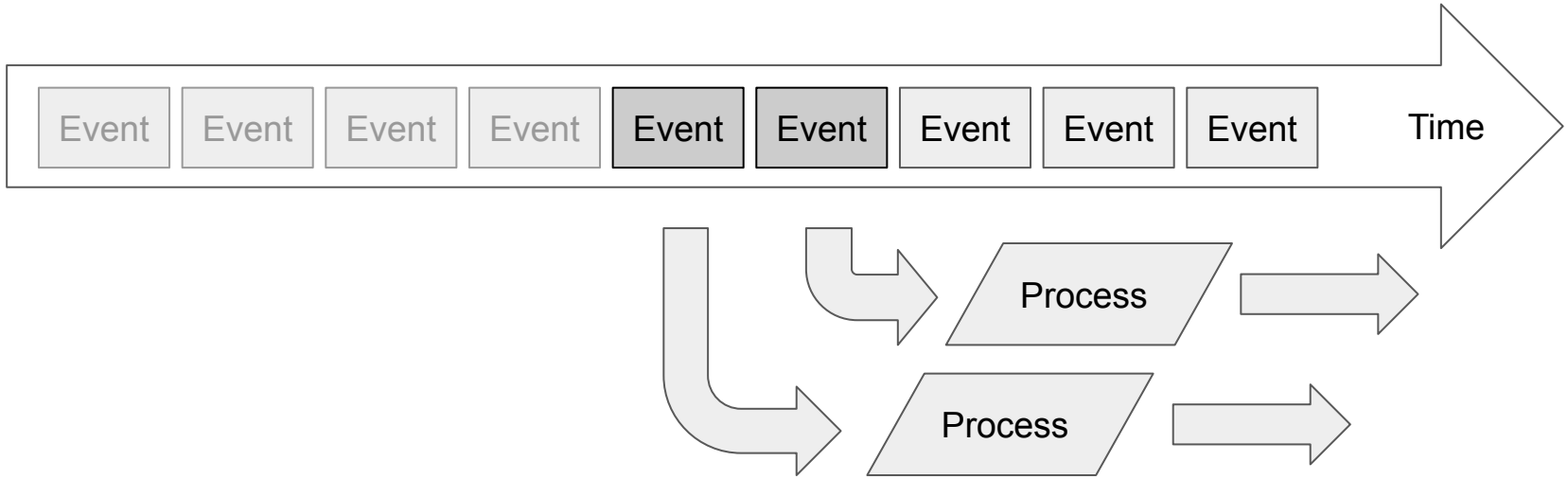


# Пособытийный подход




- Обработка одного события за раз

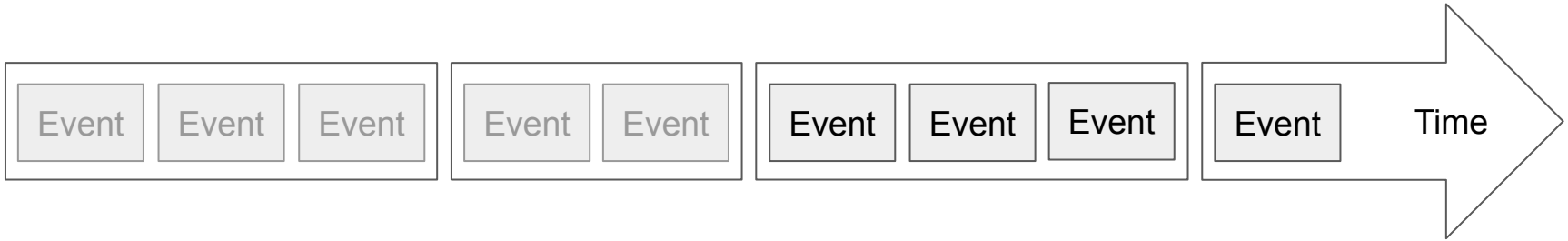
# Пособытийный подход




- Обработка одного события за раз
- События обрабатываются независимо друг от друга
- Задержка ~10ms

BIGDATA TEAM

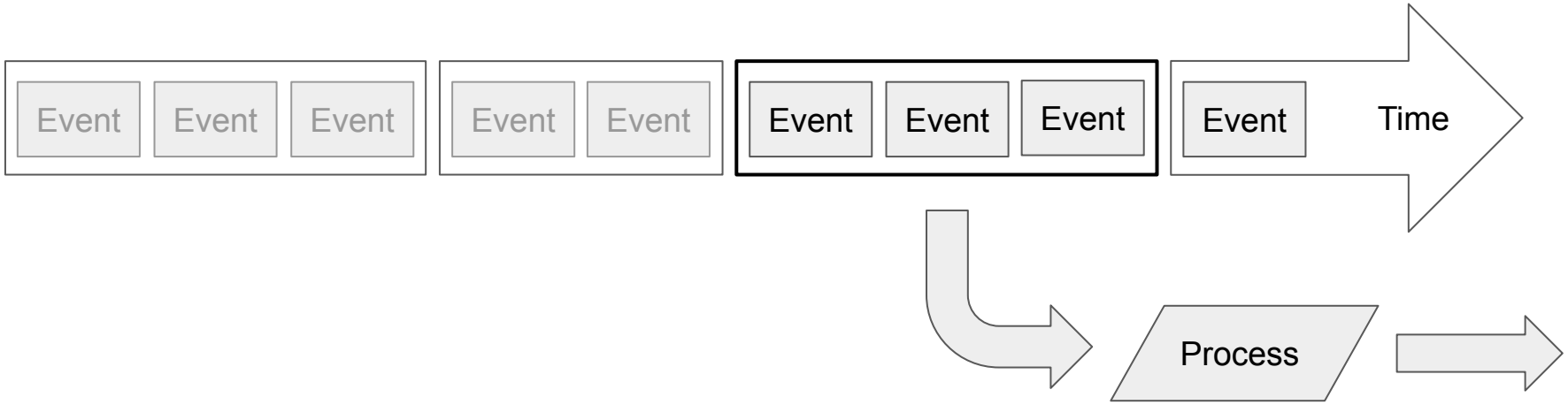
# Микробатчевый подход




- Поток событий нарезается на батчи (к примеру батч формируется из событий собранных за каждые 10 секунд)

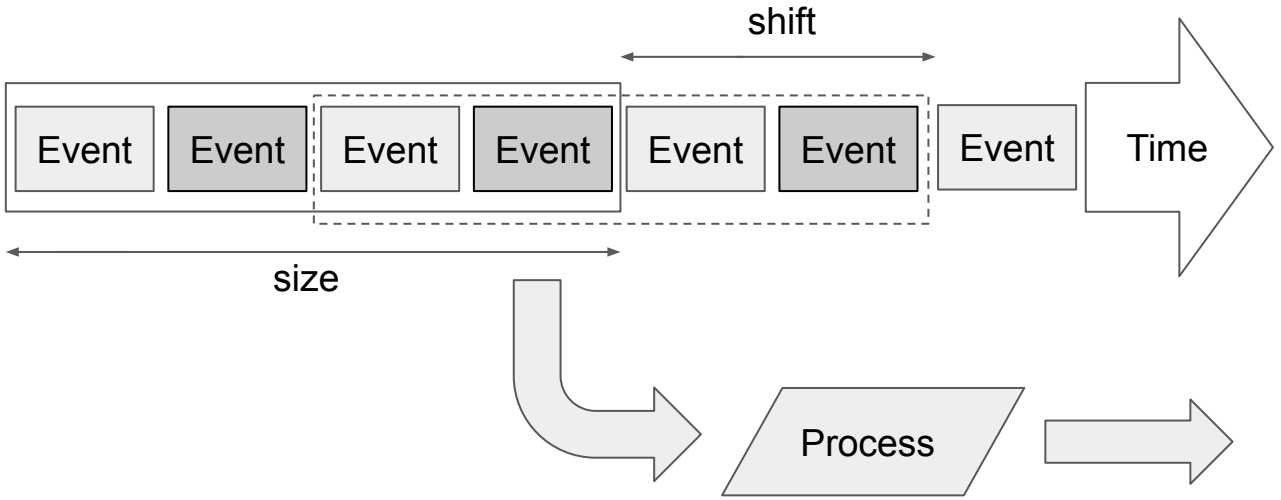
BIGDATA TEAM

# Микробатчевый подход




- Поток событий нарезается на батчи (к примеру батч формируется из событий собранных за каждые 10 секунд)
- Обычно батчи обрабатывают последовательно
- Задержка >> 1s

# Оконный подход




- Батч формируется с помощью скользящего окна
- Подвид микробатчевого подхода

BIGDATA TEAM

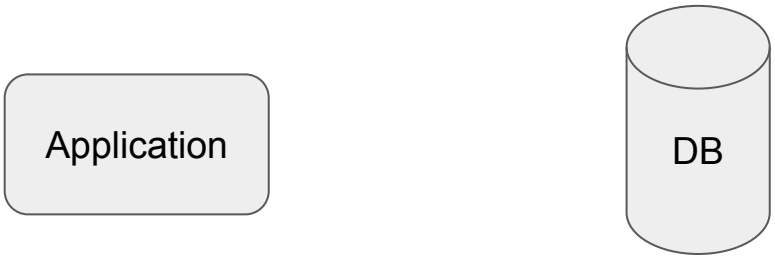
# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы

BIGDATA TEAM

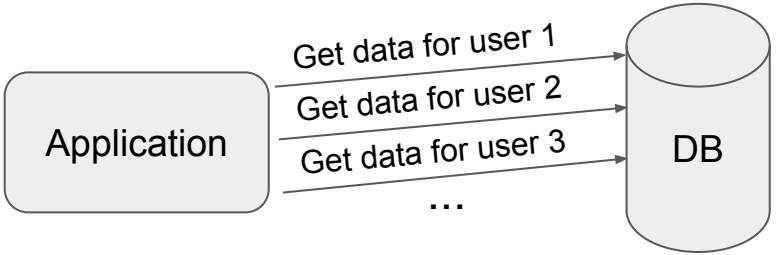
# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



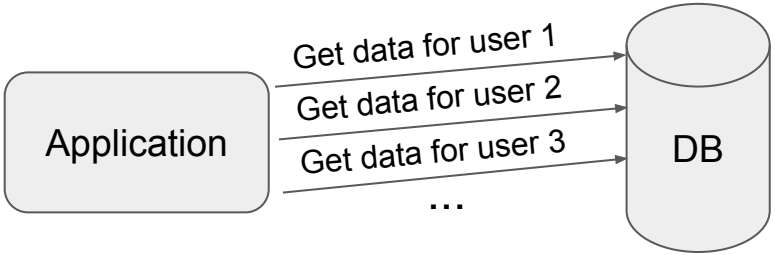
# Пособытийный vs Микробатч

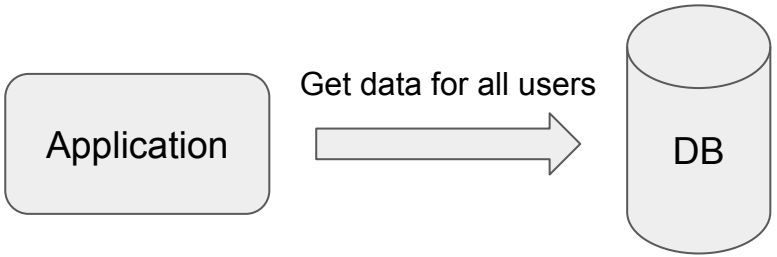
- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



# Пособытийный vs Микробатч

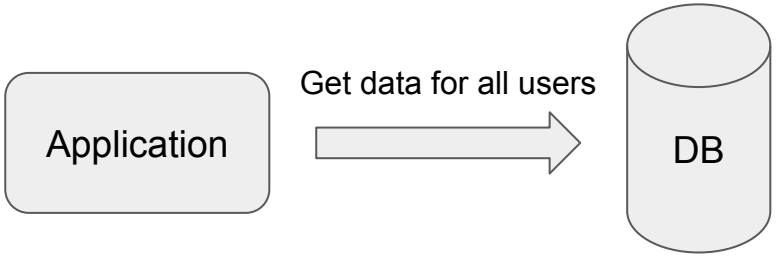
- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



# Пособытийный vs Микробатч

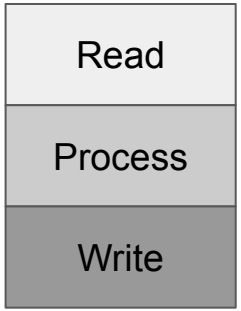
- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы

BIGDATA TEAM

# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



BIGDATA TEAM

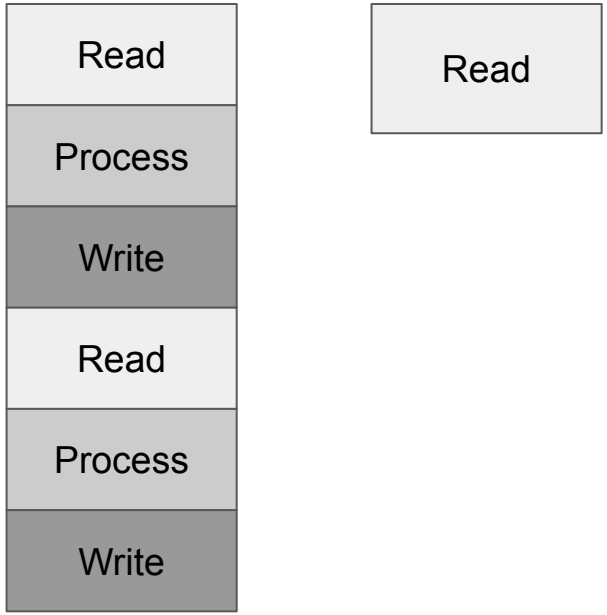
# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы

BIGDATA TEAM

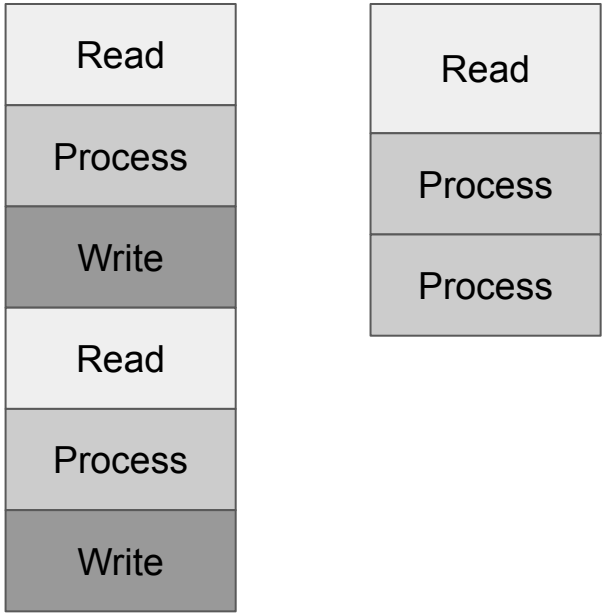
# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



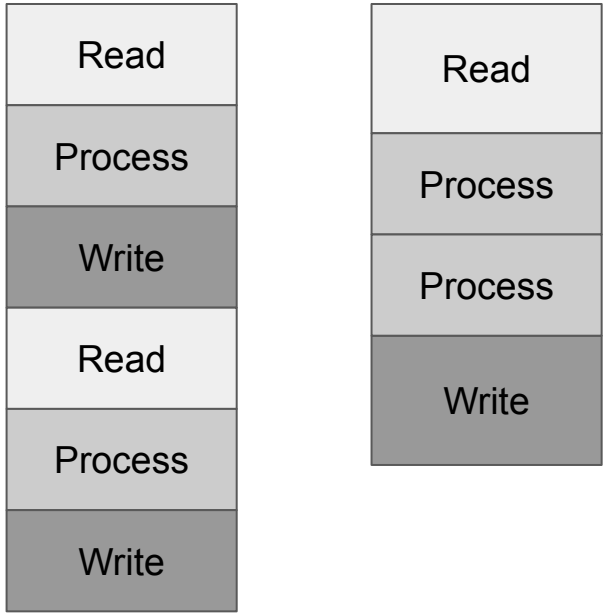
# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



# Пособытийный vs Микробатч

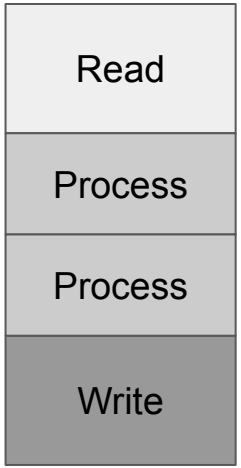
- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы



BIGDATA TEAM

# Пособытийный vs Микробатч

- Пособытийный подход позволяет достичь наименьшего лага
- Микробатч позволяет сэкономить ресурсы
- Микробатч позволяет обработать больше данных на тех же ресурсах в сравнении с пособытийным

BIGDATA TEAM

# Пропускная способность VS Задержка

- В реальной жизни ресурсы всегда ограничены
- Большие данные требуют огромной пропускной способности => в большинстве случаев используем микробатч
- В каждом отдельном случае вы должны выбирать решение согласно задаче

# Пропускная способность VS Задержка