

PGDAY
RUSSIA 17



АНТИ-введение в Big Data

Владимир Красильщик, Яндекс
Санкт-Петербург, Россия

Что за чувак переключает слайды?

Что за чувак переключает слайды?

> Альма-матер: СПбГЭТУ “ЛЭТИ”

Что за чувак переключает слайды?

- > Альма-матер: СПбГЭТУ “ЛЭТИ”
- > До Luxoft: таможня РФ и Беларуси, Continental, Daimler-Chrysler, школы Баварии и Баден-Вюртенберга, T-Mobile

Что за чувак переключает слайды?

- > Альма-матер: СПбГЭТУ “ЛЭТИ”
- > До Luxoft: таможня РФ и Беларуси, Continental, Daimler-Chrysler, школы Баварии и Баден-Вюртенберга, T-Mobile
- > Luxoft: крупный инвестиционный банк

Что за чувак переключает слайды?

- > Альма-матер: СПбГЭТУ “ЛЭТИ”
- > До Luxoft: таможня РФ и Беларуси, Continental, Daimler-Chrysler, школы Баварии и Баден-Вюртенберга, T-Mobile
- > Luxoft: крупный инвестиционный банк
- > Яндекс

Что за чувак переключает слайды?

- > Альма-матер: СПбГЭТУ “ЛЭТИ”
- > До Luxoft: таможня РФ и Беларуси, Continental, Daimler-Chrysler, школы Баварии и Баден-Вюртенберга, T-Mobile
- > Luxoft: крупный инвестиционный банк
- > Яндекс

Какова цель выступления?



Big Data это ...

Big Data это ...

> Big Data это когда петабайты данных

Big Data это ...

- > Big Data это когда петабайты данных
- > Big Data это highload

Big Data это ...

- > Big Data это когда петабайты данных
- > Big Data это highload
- > Big Data это hadoop

Big Data это ...

- > Big Data это когда петабайты данных
- > Big Data это highload
- > Big Data это hadoop
- > Big Data это треш и угар

Big Data это ...

big data это



big data это

big data **что** это такое

big data это **сколько**

big data — это **философия**

About 61,500,000 results (0.32 seconds)

[Аналитика Big Data - Microsoft.com](#)

Ad www.microsoft.com/ ▼

Получайте быстрые и ценные инсайты используя надежное BI решение!

Другими словами, **Big Data** – это проблема. Проблема хранения и обработки гигантских объемов данных. С другой стороны, обработка больших объемов информации – **это** только часть «айсберга».



Big Data (биг дата) — обработка больших объемов данных

www.dis-group.ru/solutions/data_management/big_data/

Big Data это ...

big data это



big data это

big data **что это такое**

big data это **сколько**

big data — это **философия**

About 1,000,000 results (0.32 seconds)

Аналитика Big Data - Microsoft.com

Ad www.microsoft.com/ ▼

Получайте быстрые и ценные инсайты используя надежное BI решение!

Другими словами, **Big Data** – это проблема. Проблема хранения и обработки гигантских объемов данных. С другой стороны, обработка больших объемов информации – **это** только часть «айсберга».



Big Data (биг дата) — обработка больших объемов данных

www.dis-group.ru/solutions/data_management/big_data/

Big Data это ...

big data это



big data это

big data **что это такое**

big data это **сколько**

big data — это **философия**

About 61,000,000 results (0.32 seconds)

Аналитика Big Data - Microsoft.com

Ad www.microsoft.com/ ▼

Получайте быстрые и ценные инсайты используя надежное BI решение!

Другими словами, **Big Data – это проблема.** Проблема хранения и обработки гигантских объемов данных. С другой стороны, обработка больших объемов информации – **это** только часть «айсберга».



Big Data (биг дата) — обработка больших объемов данных

www.dis-group.ru/solutions/data_management/big_data/

Big Data это ...

Карты Маркет Новости Переводчик Картинки Видео Музыка ещё

Яндекс

big data что

Найти

- big data что это такое
- big data курсы
- big data dangerous
- big data это
- big data что это
- big data анализ
- big data реферат
- big data 2017
- big data книги
- big data обучение

ТОМ КРУЗ
СМОТРИТЕ ТРЕЙЛЕР

Погода ☁ +20
День +20, вечер +16

Посещаемое

Маркет — шпатлёвка до 3000

Авто.ру — где починить машину

Недвижимость — жилой район

10:20 Лесник НТВ

10:30 Дом-2. Остров любви ТНТ

10:55 Модный приговор Первый

Афиша

Нелюбовь премьера в чт

Пираты Карибского моря... приключения

исания

обусы

Big Data это ...

Карты Маркет Новости Переводчик Картинки Видео Музыка ещё

Яндекс

big data что

big data что это такое

big data курсы

big data dangerous

big data это

big data что это

big data анализ

big data реферат

big data 2017

big data книги

big data обучение

Найти

ТОМ КРУЗ
СМОТРИТЕ ТРЕЙЛЕР

Погода ☁ +20
День +20, вечер +16

Посещаемое

Маркет — шпатлёвка до 3000

Авто.ру — где починить машину

Недвижимость — жилой район

10:20 Лесник НТВ

10:30 Дом-2. Остров любви ТНТ

10:55 Модный приговор Первый

Афиша

Нелюбовь премьера в чт

Пираты Карибского моря... приключения

исания

обусы

**Сижу жду когда мне расскажут
про Big Data**



Что такое Big Data

Big Data это набор мер и технологий, направленный на монетизацию имеющихся у компании данных



Данные

Технологии

Аналитика

Продукты и маркетинг

Процессы

Что такое Big Data

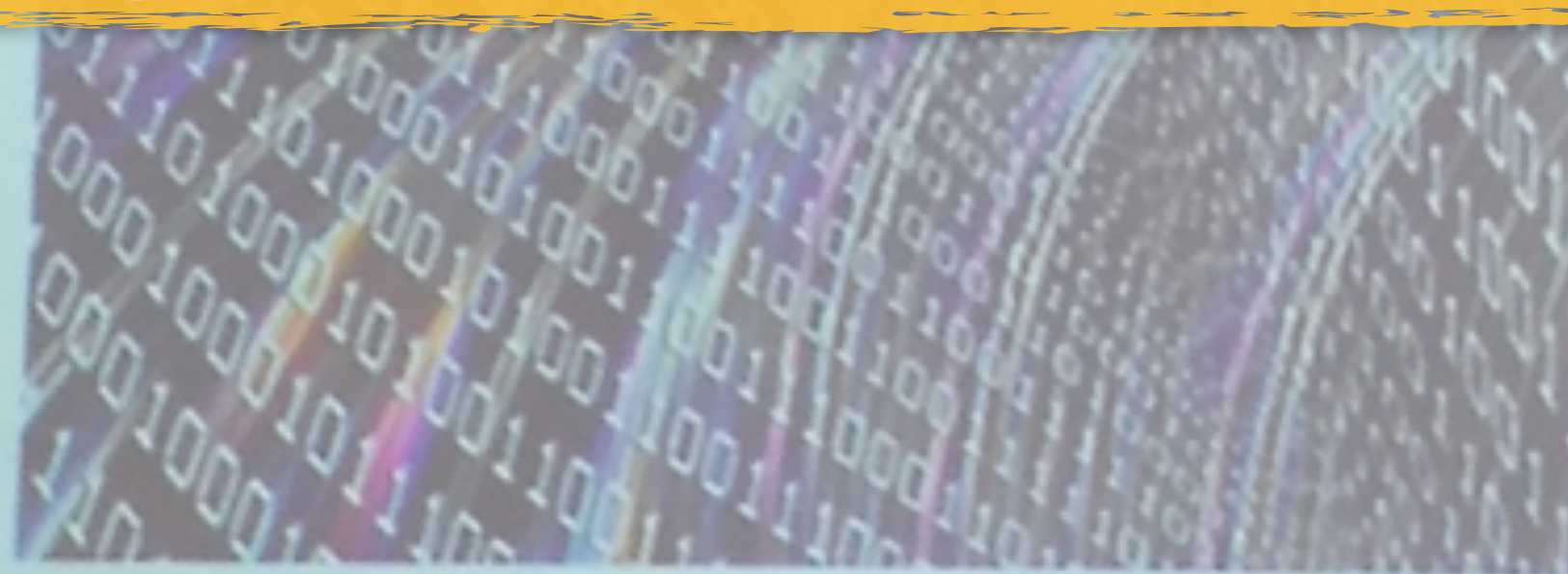
Big Data это набор мер и технологий, направленный на монетизацию имеющихся у компании данных

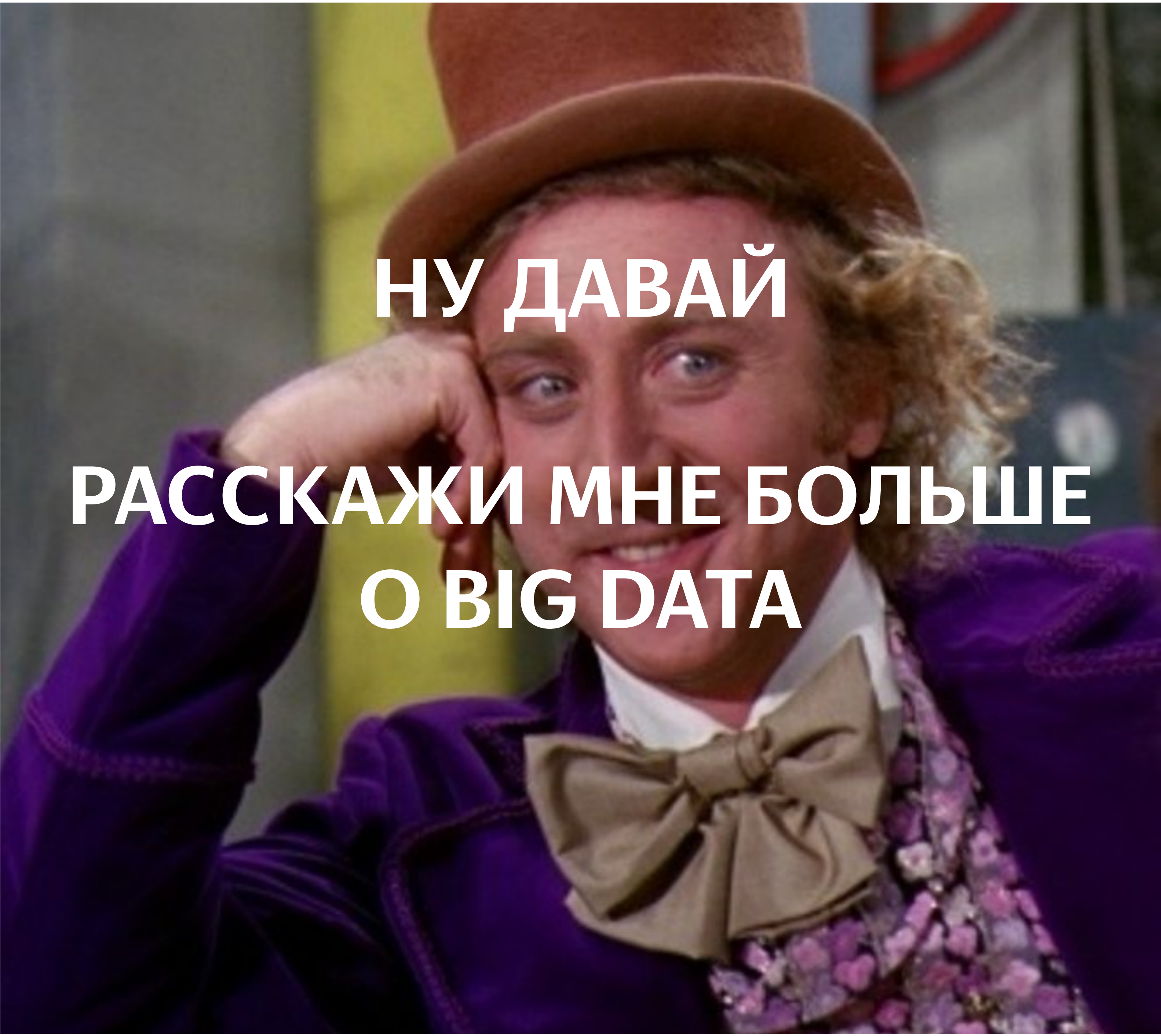
Данные

Технологии

Big Data это набор мер и технологий, направленный на монетизацию имеющихся у компании данных

Процессы



A close-up photograph of Gene Wilder, dressed in a purple velvet suit, a brown top hat, and a large tan bow tie. He is smiling and leaning his head on his right hand. The background is slightly blurred, showing a yellow and blue structure.

**НУ ДАВАЙ
РАССКАЖИ МНЕ БОЛЬШЕ
О BIG DATA**

Цель выступления - прочувствовать
что такое Big Data



Почему доклад называется
«анти-введение»?



Вопрос

Вопрос

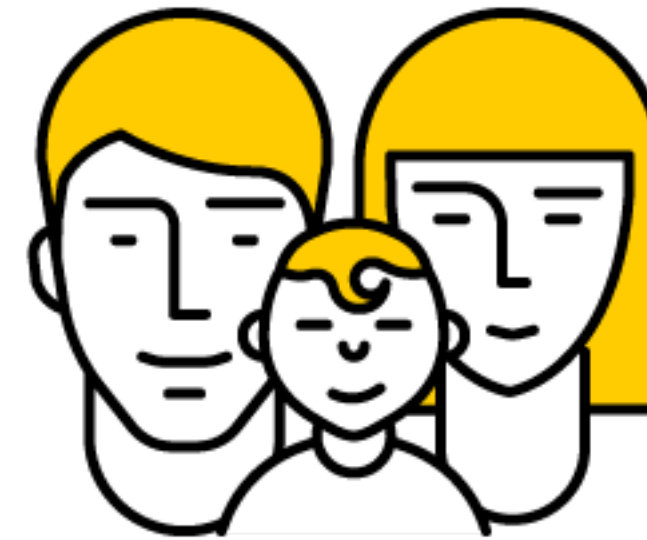
Сколько транзакций в день обрабатывается в крупном инвестиционном банке?

Транзакции

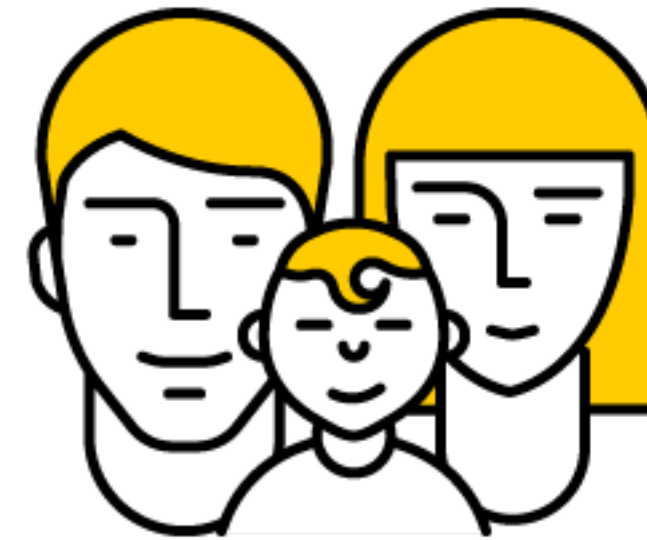
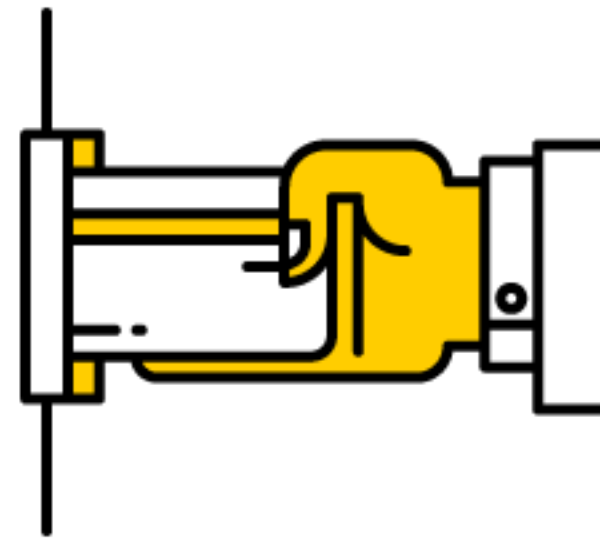
Транзакции



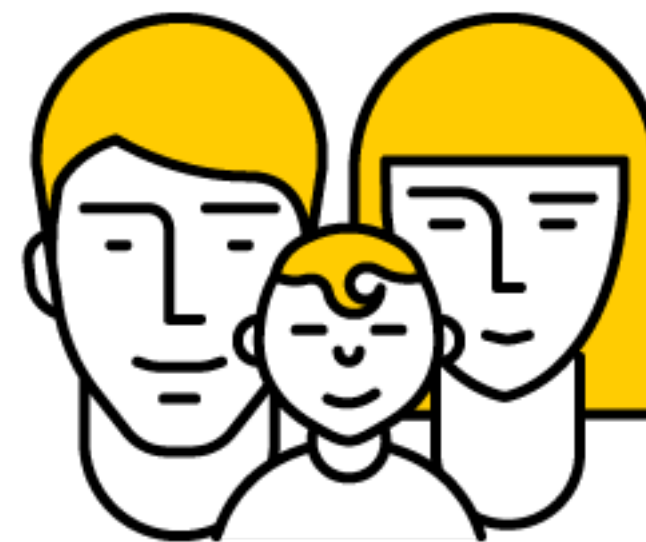
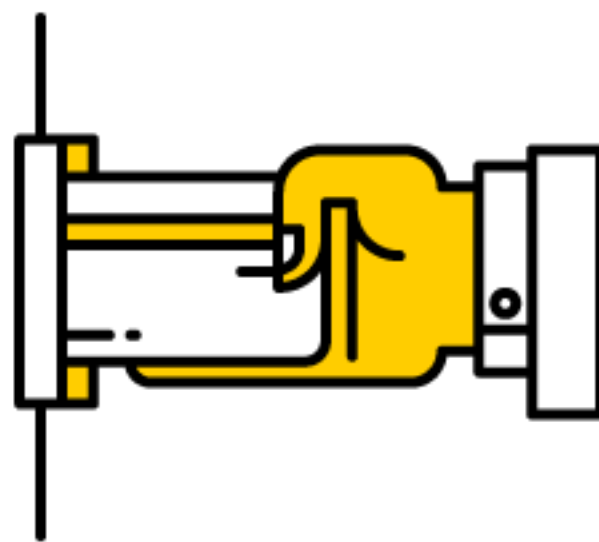
Транзакции



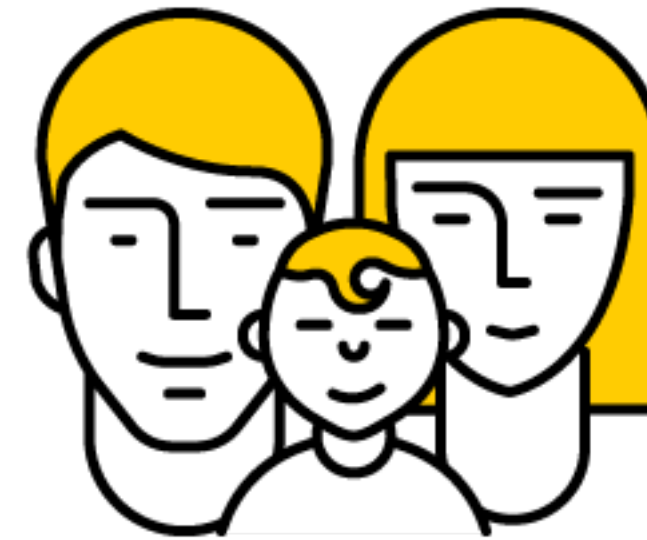
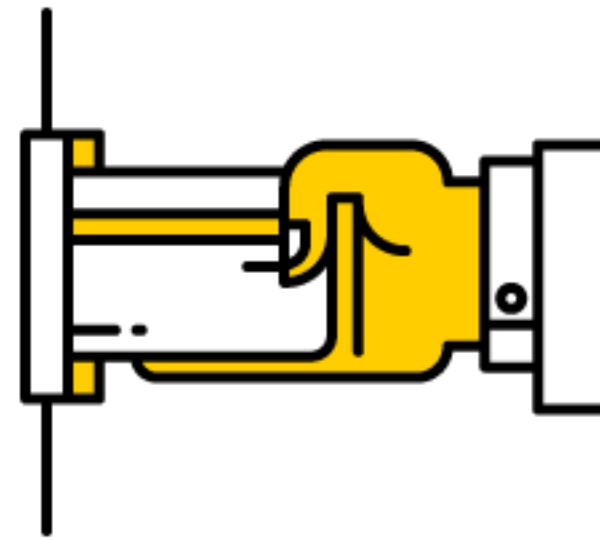
Транзакции



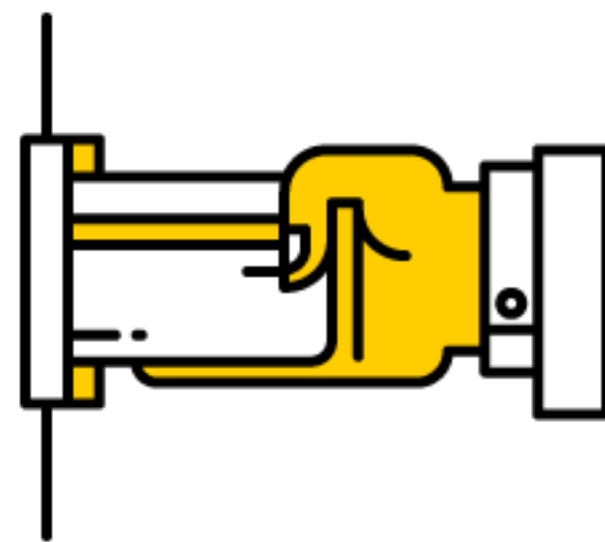
Транзакции



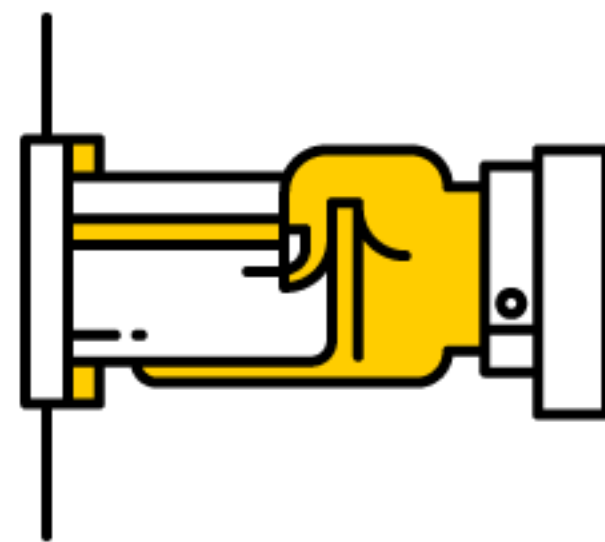
Транзакции



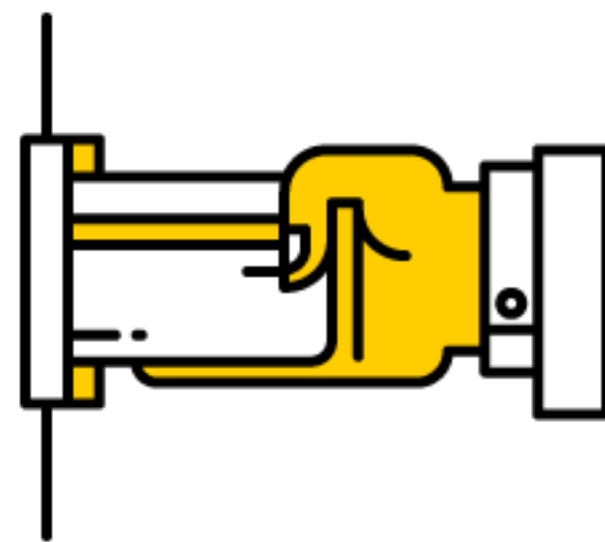
Транзакции



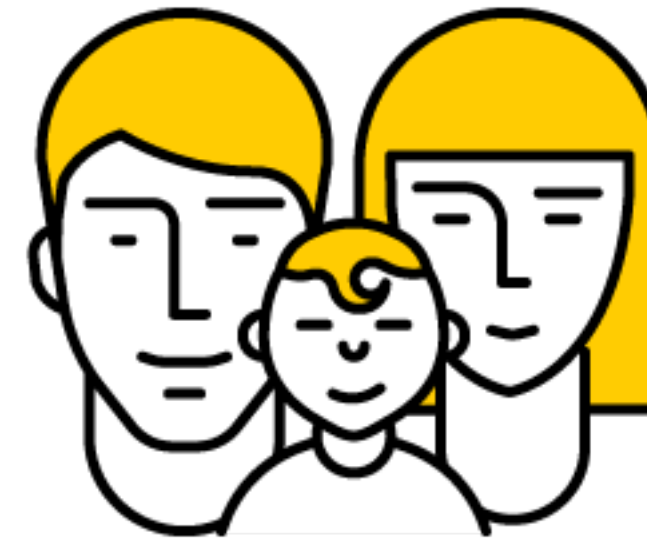
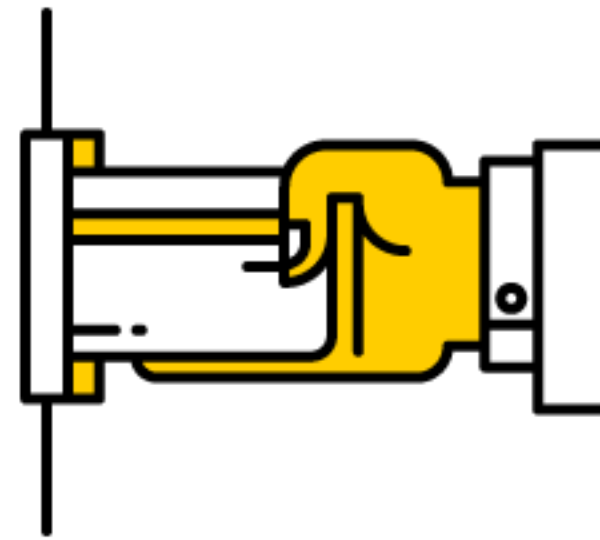
Транзакции



Транзакции



Транзакции



Вопрос

Вопрос

Сколько транзакций в день обрабатывается
в крупном инвестиционном банке?

Вопрос

Сколько транзакций в день обрабатывается
в крупном инвестиционном банке?

1.1M

Вопрос

Сколько транзакций в день обрабатывается
в крупном инвестиционном банке?

1. 1M

2. 100M

Вопрос

Сколько транзакций в день обрабатывается
в крупном инвестиционном банке?

- 1. 1M
- 2. 100M
- 3. 500M

Вопрос

Сколько транзакций в день обрабатывается
в крупном инвестиционном банке?

1. 1M

2. 100M

3. 500M

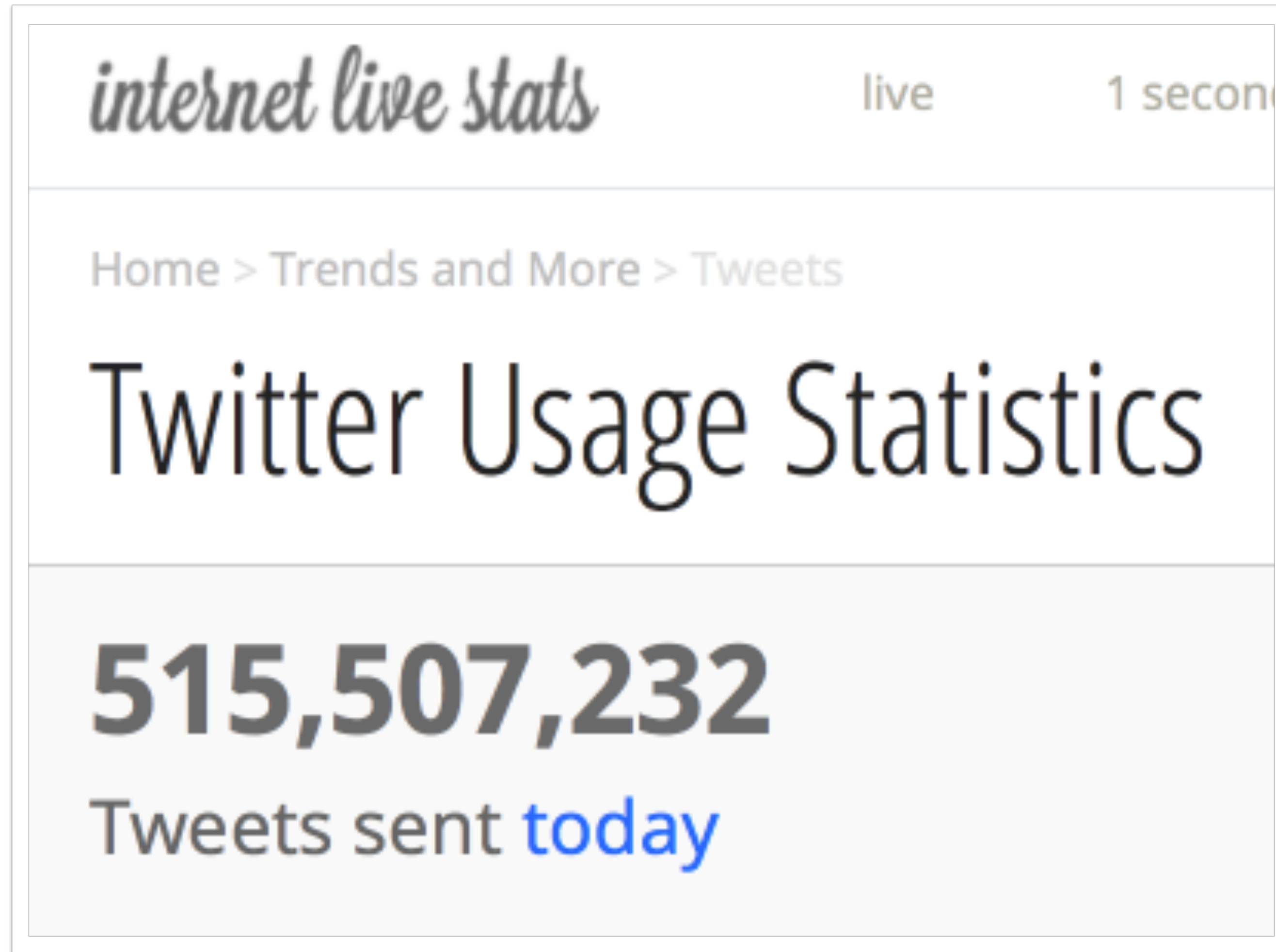
до 500M



twitter



<http://www.internetlivestats.com/twitter-statistics/>



The screenshot shows the 'internet live stats' website with a 'live' indicator and a '1 second' update timer. The breadcrumb trail is 'Home > Trends and More > Tweets'. The main title is 'Twitter Usage Statistics'. The primary statistic is '515,507,232' in large bold text, with the label 'Tweets sent today' below it, where 'today' is highlighted in blue.

internet live stats live 1 second

Home > Trends and More > Tweets

Twitter Usage Statistics

515,507,232
Tweets sent **today**

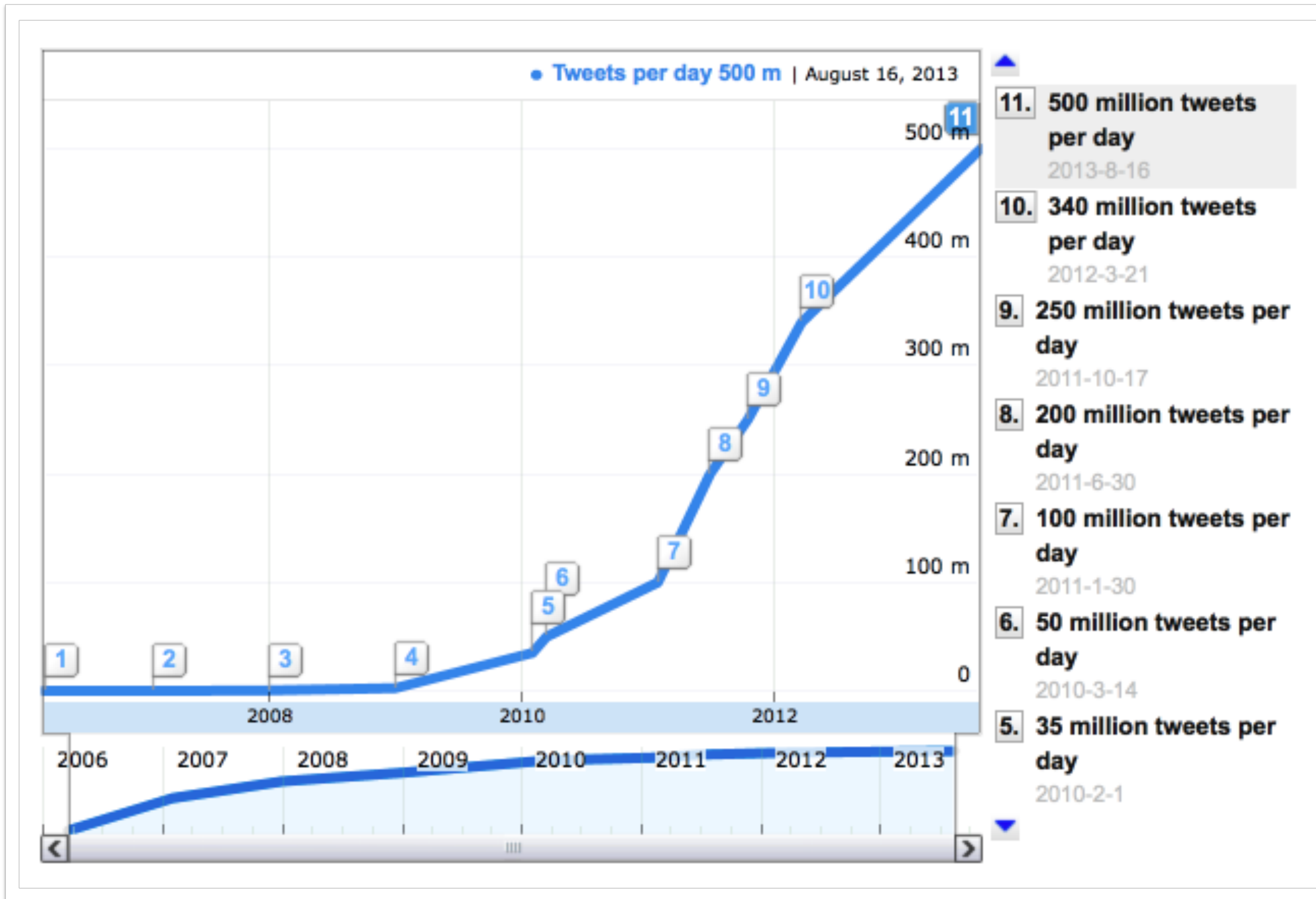
29-05-2017
23:15

<http://www.internetlivestats.com/twitter-statistics/>

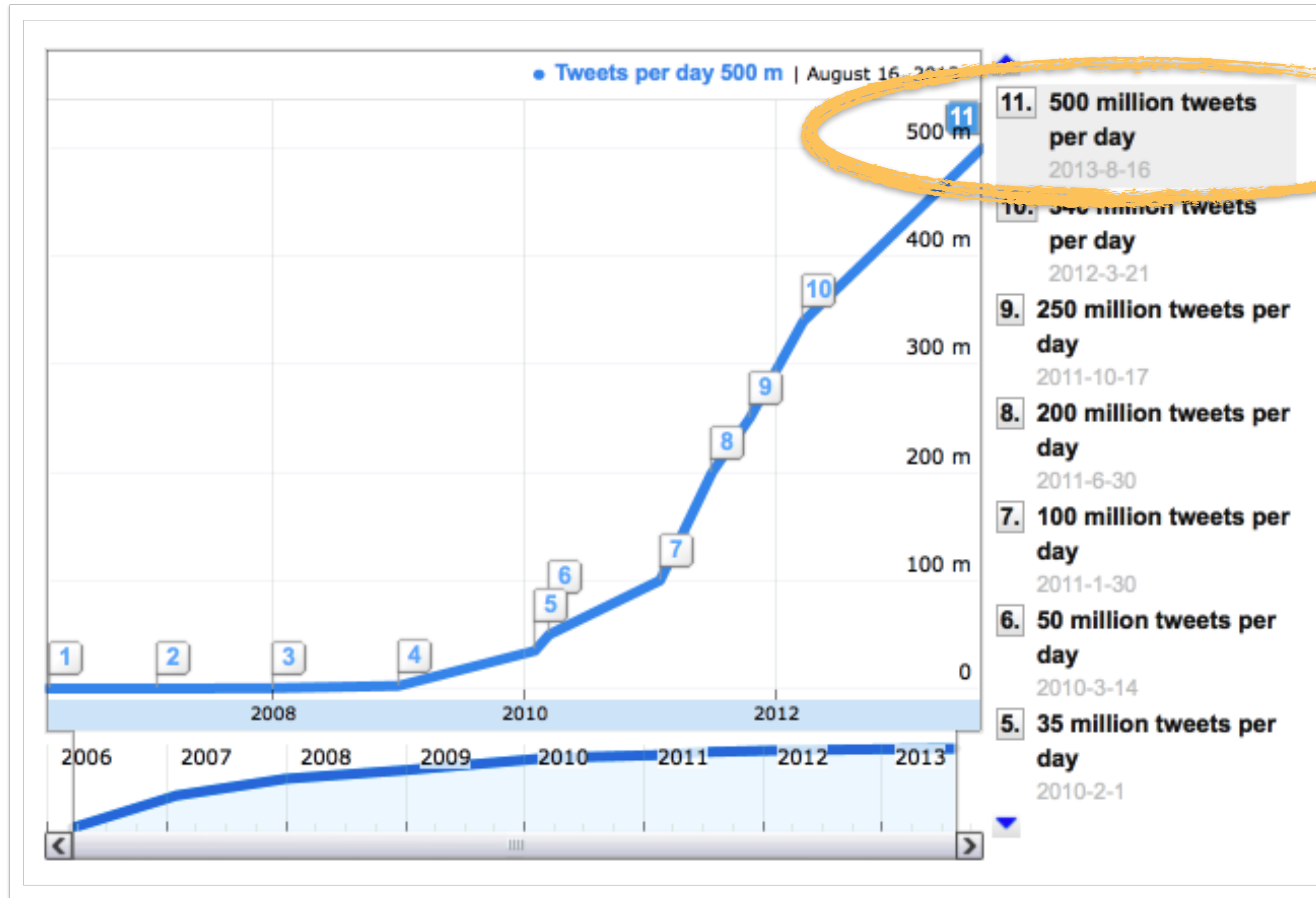
29-05-2017
23:15



<http://www.internetlivestats.com/twitter-statistics/>



<http://www.internetlivestats.com/twitter-statistics/>



Что такое ...

#hashtag

умные-решения.рф

умные-решения.рф

Какие сейчас* самые популярные *хештеги*?

умные-решения.рф

Какие сейчас* самые популярные *хештеги*?

Сейчас* - на настоящий момент с начала времён

<https://twitter.com/tophashtags>

Tweets

Tweets & replies



Top Hashtags @TopHashtags · 13m

In Australia/6h: 1.[#auspol](#) 2.[#houserules](#) 3.[#monacogp](#) 4.[#masterchefau](#)
5.[#thevoiceau](#)



Top Hashtags @TopHashtags · 25m

In Deutschland/6h: 1.[#immerwiederbvb](#) 2.[#rpc](#) 3.[#dfbpokal](#) 4.[#afd](#)
5.[#roleplayconvention](#)

Translate from German



Top Hashtags @TopHashtags · 31m

En España/6h: 1.[#halamadrid](#) 2.[#felizdomingo](#) 3.[#felizlunes](#) 4.[#33ligas](#)
5.[#madrid](#)

Translate from Spanish



Top Hashtags @TopHashtags · 2h

En France/6h: 1.[#fifau20wc](#) 2.[#fetedesmeres](#) 3.[#concours](#) 4.[#rt](#)
5.[#fêtedesmères](#)

Translate from French

29-05-2017
12:31

<https://twitter.com/tophashtags>

Tweets

Tweets & replies

 **Top Hashtags** @TopHashtags · 13m
In Australia/6h: 1.#auspol 2.#houserules 3.#monacogp 4.#masterchefau
5.#thevoiceau

 **Top Hashtags** @TopHashtags · 25m
In Deutschland/6h: 1.#immerwiederbvb 2.#rpc 3.#dfbpokal 4.#afd
5.#roleplayconvention

 Translate from German

 **Top Hashtags** @TopHashtags · 31m
En España/6h: 1.#halamadrid 2.#felizdomingo 3.#felizlunes 4.#33ligas
5.#madrid

 Translate from Spanish

 **Top Hashtags** @TopHashtags · 2h
En France/6h: 1.#fifau20wc 2.#fetedesmeres 3.#concours 4.#rt
5.#fêtedesmères

 Translate from French

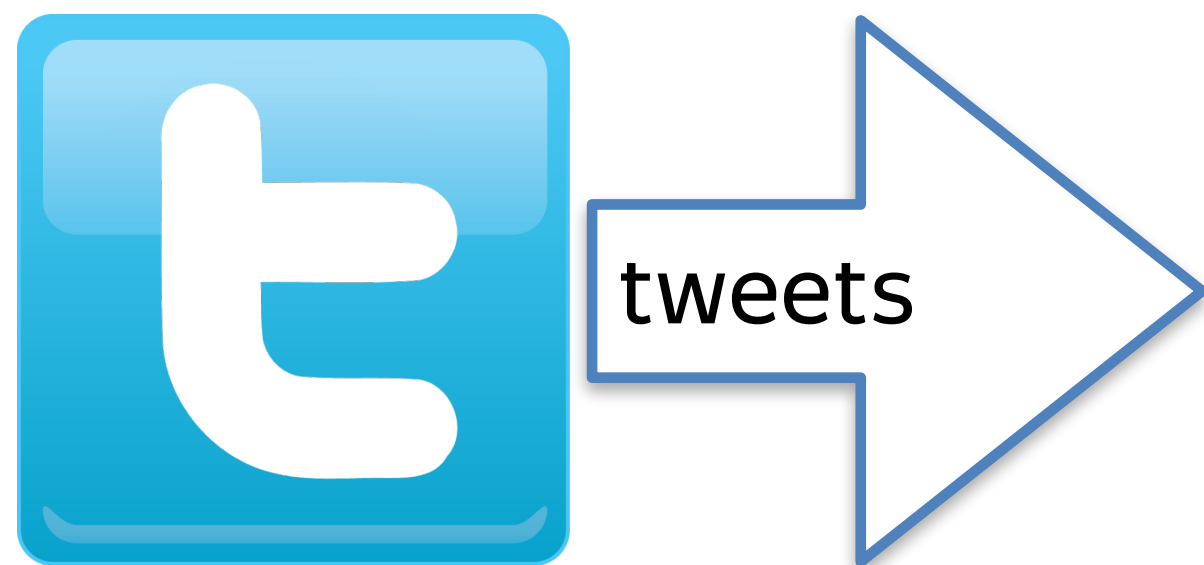
29-05-2017
12:31

Инкрементальное решение

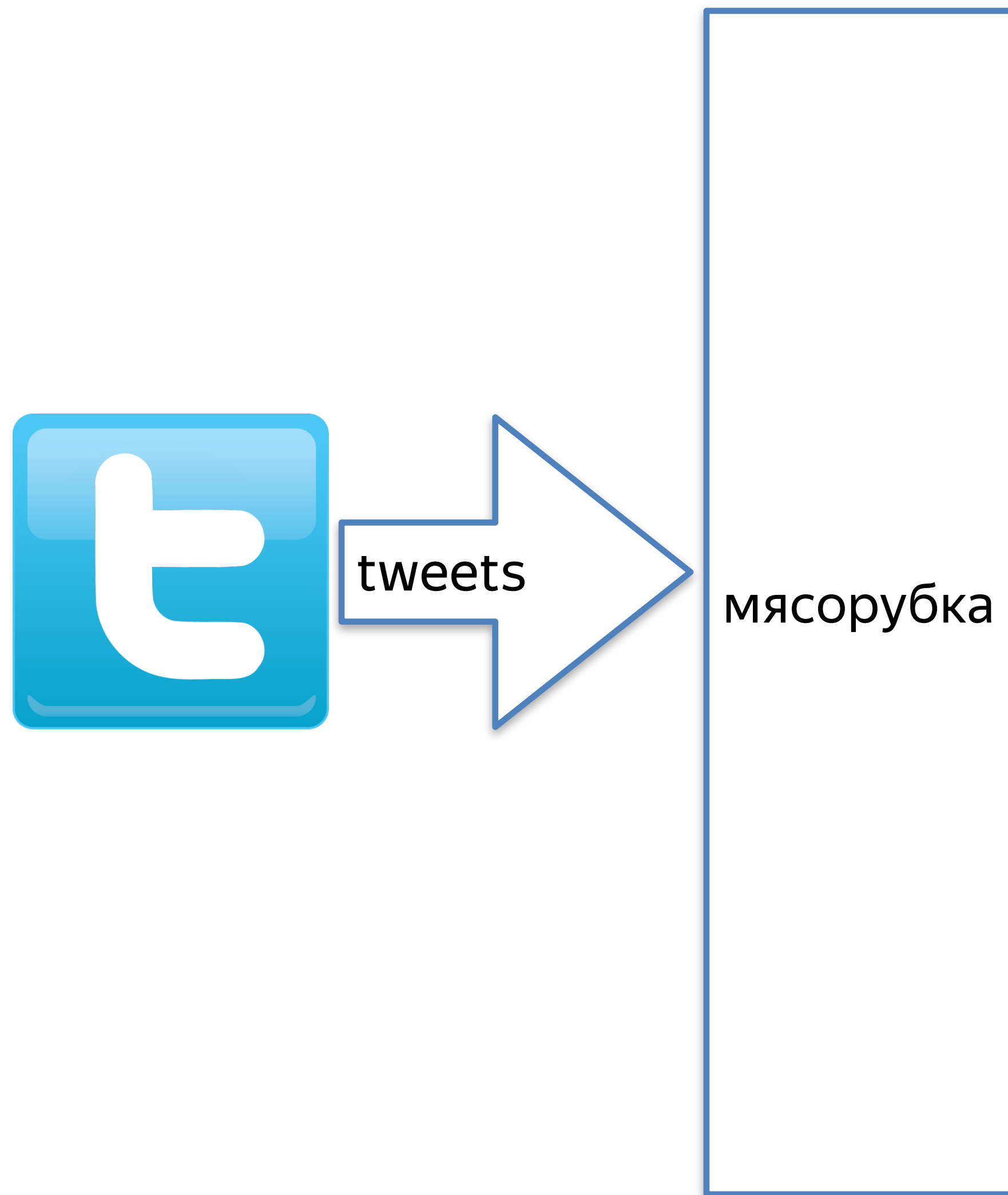
Инкрементальное решение



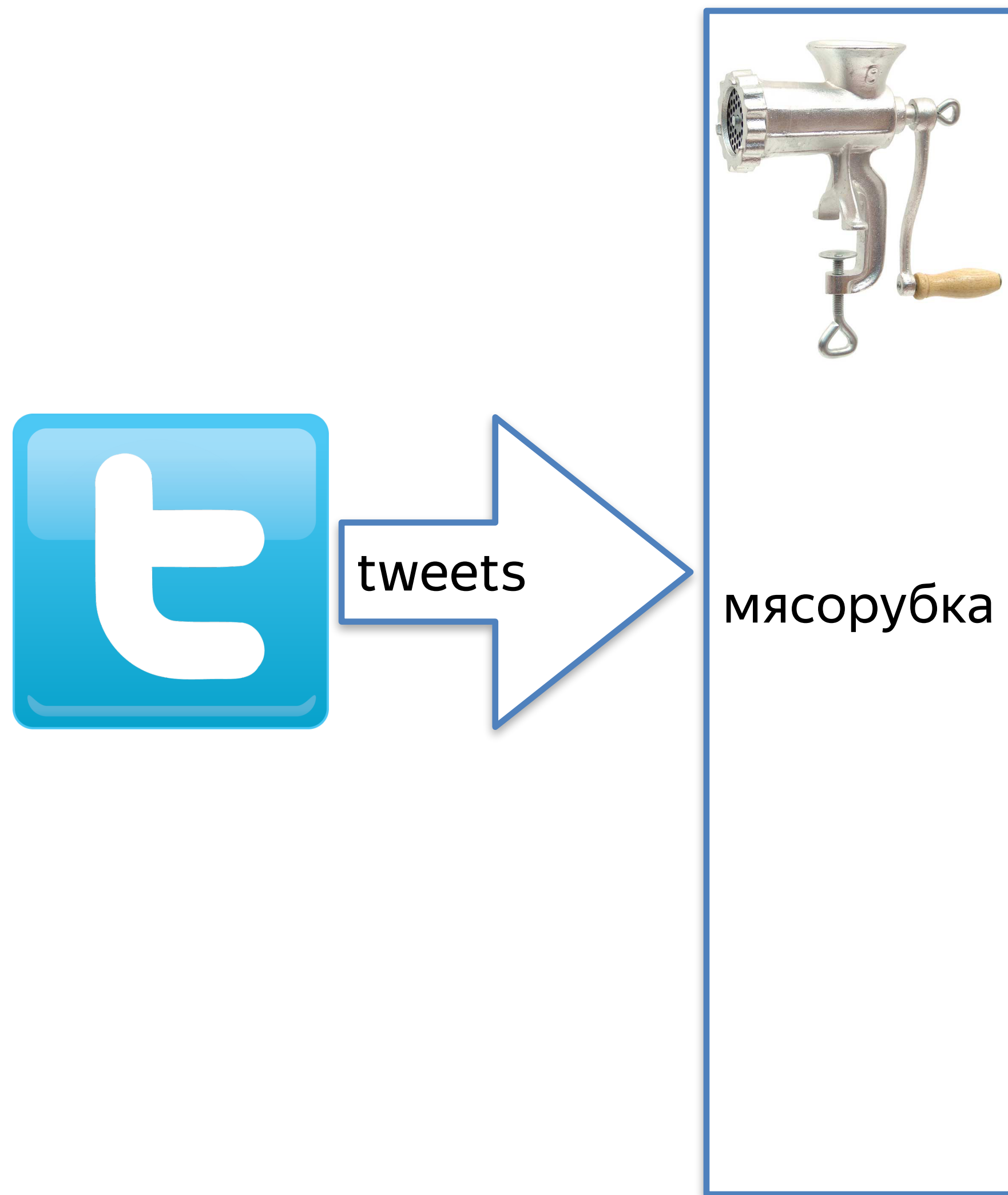
Инкрементальное решение



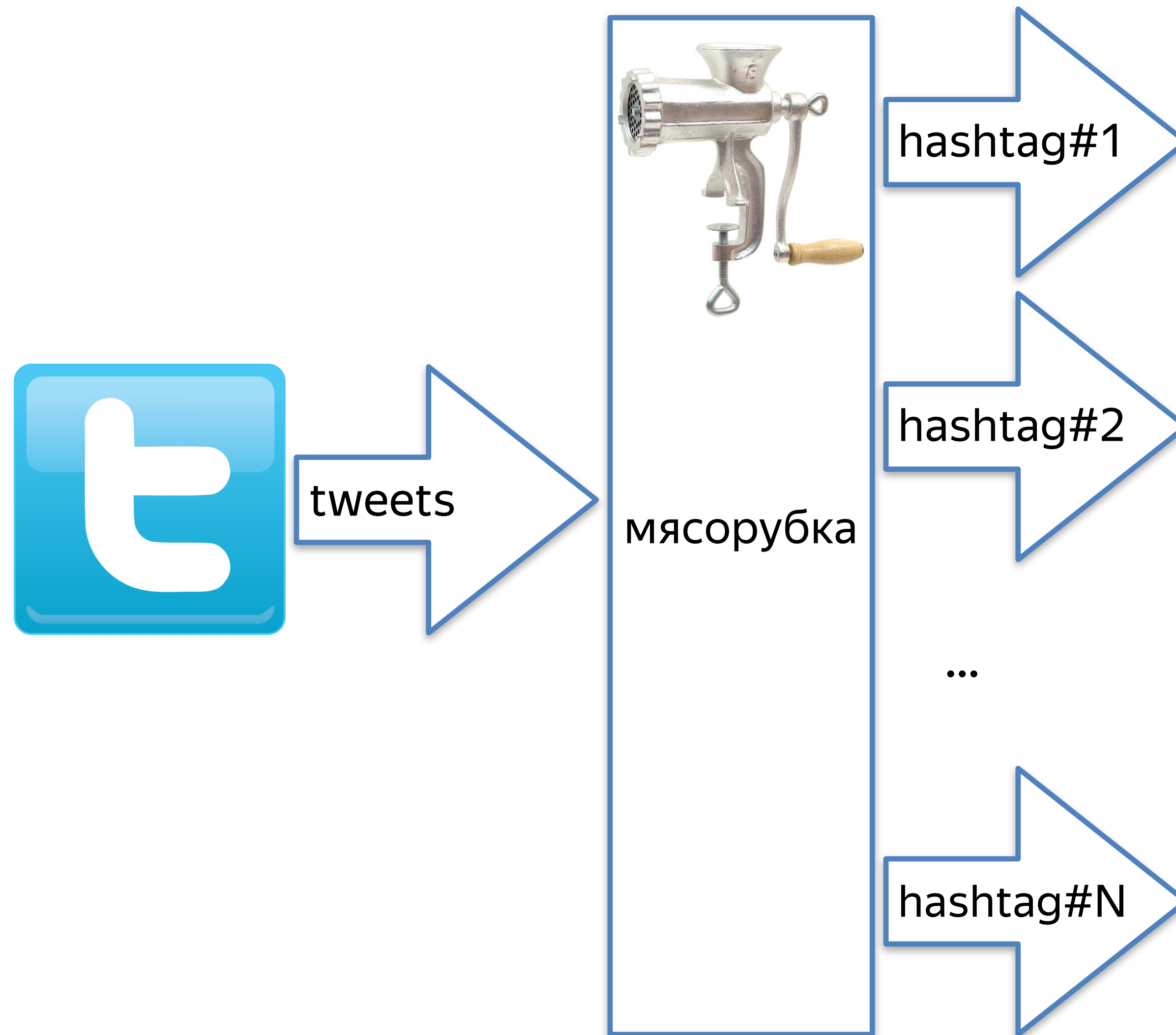
Инкрементальное решение



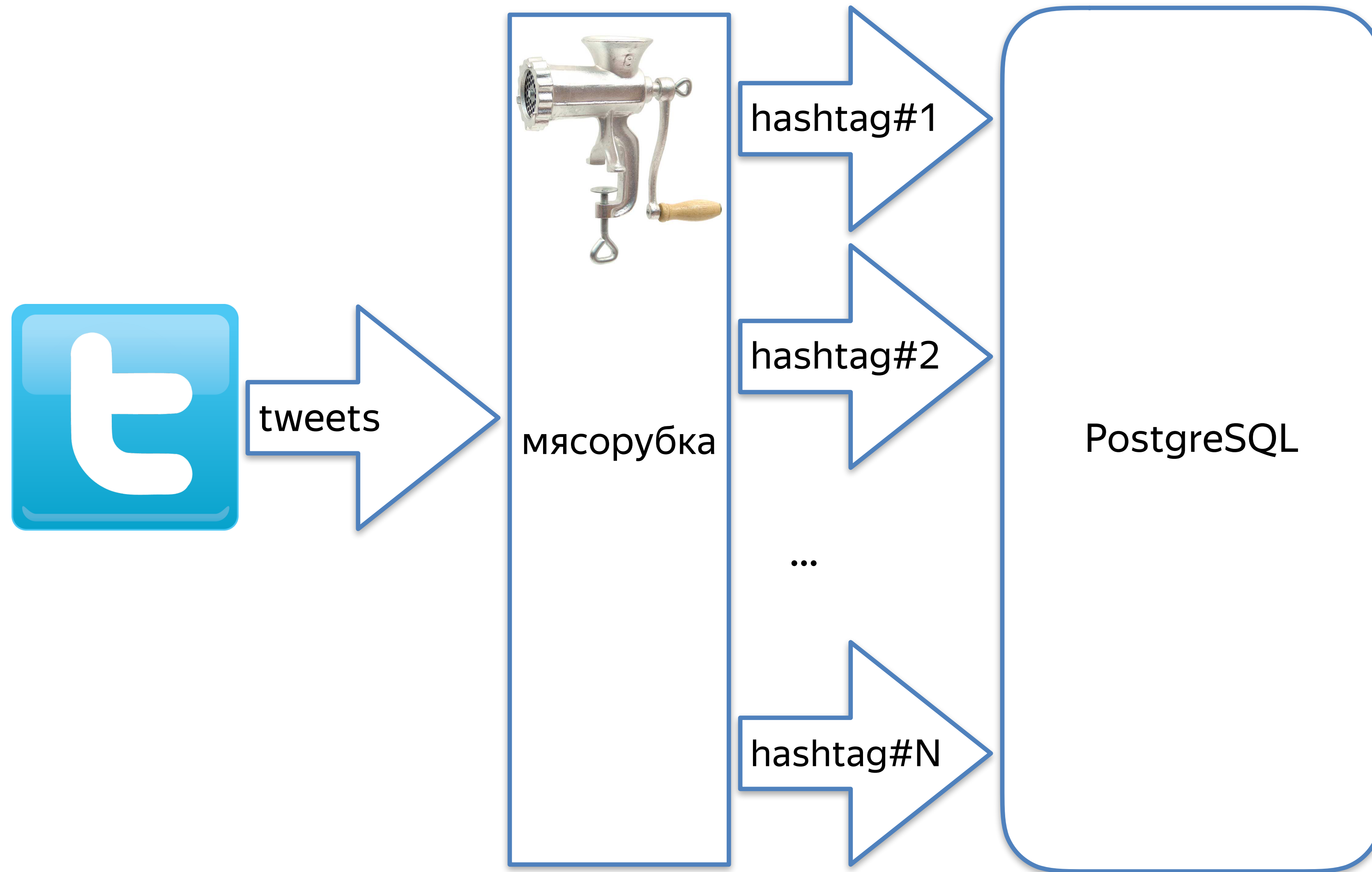
Инкрементальное решение



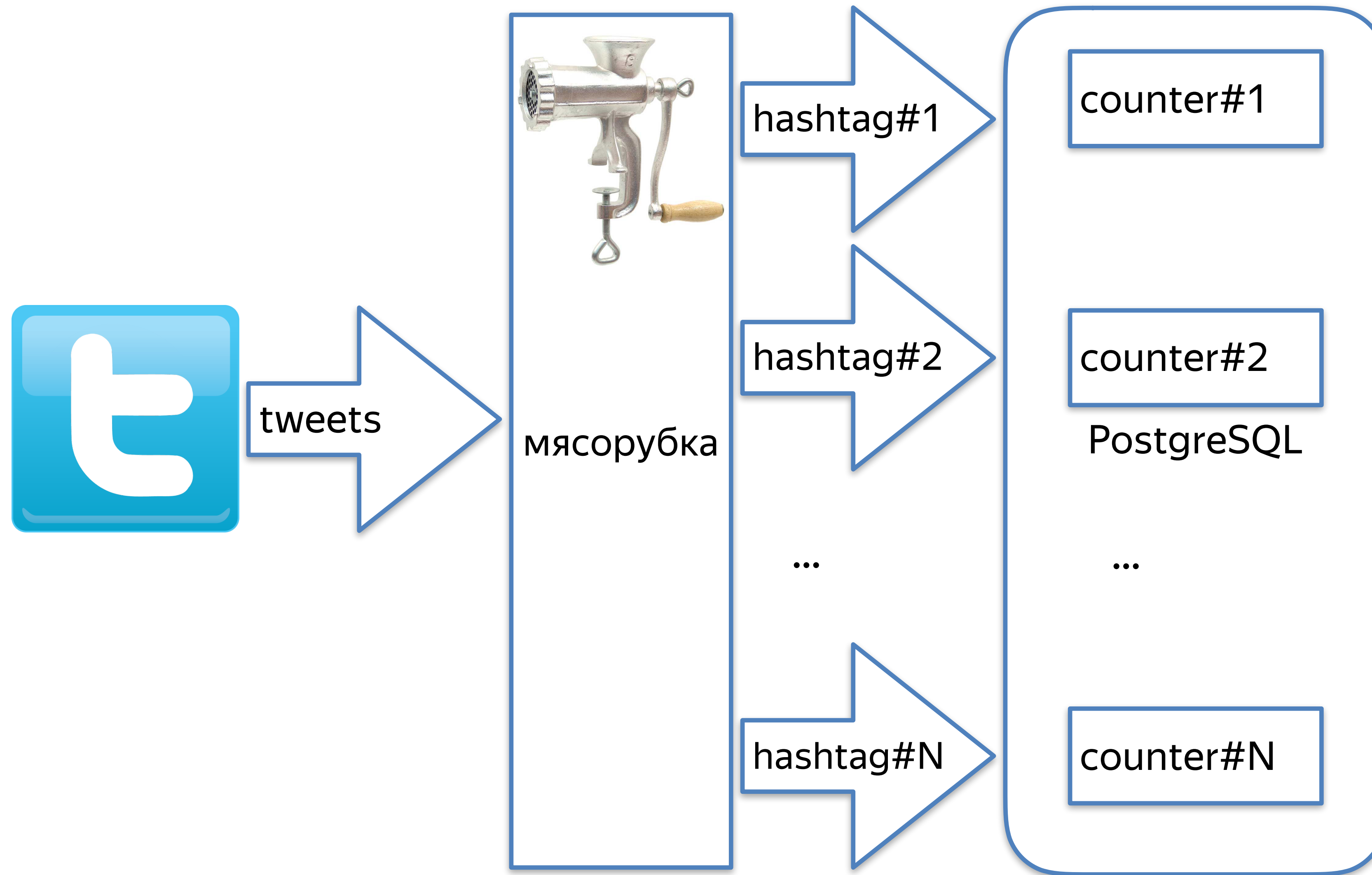
Инкрементальное решение



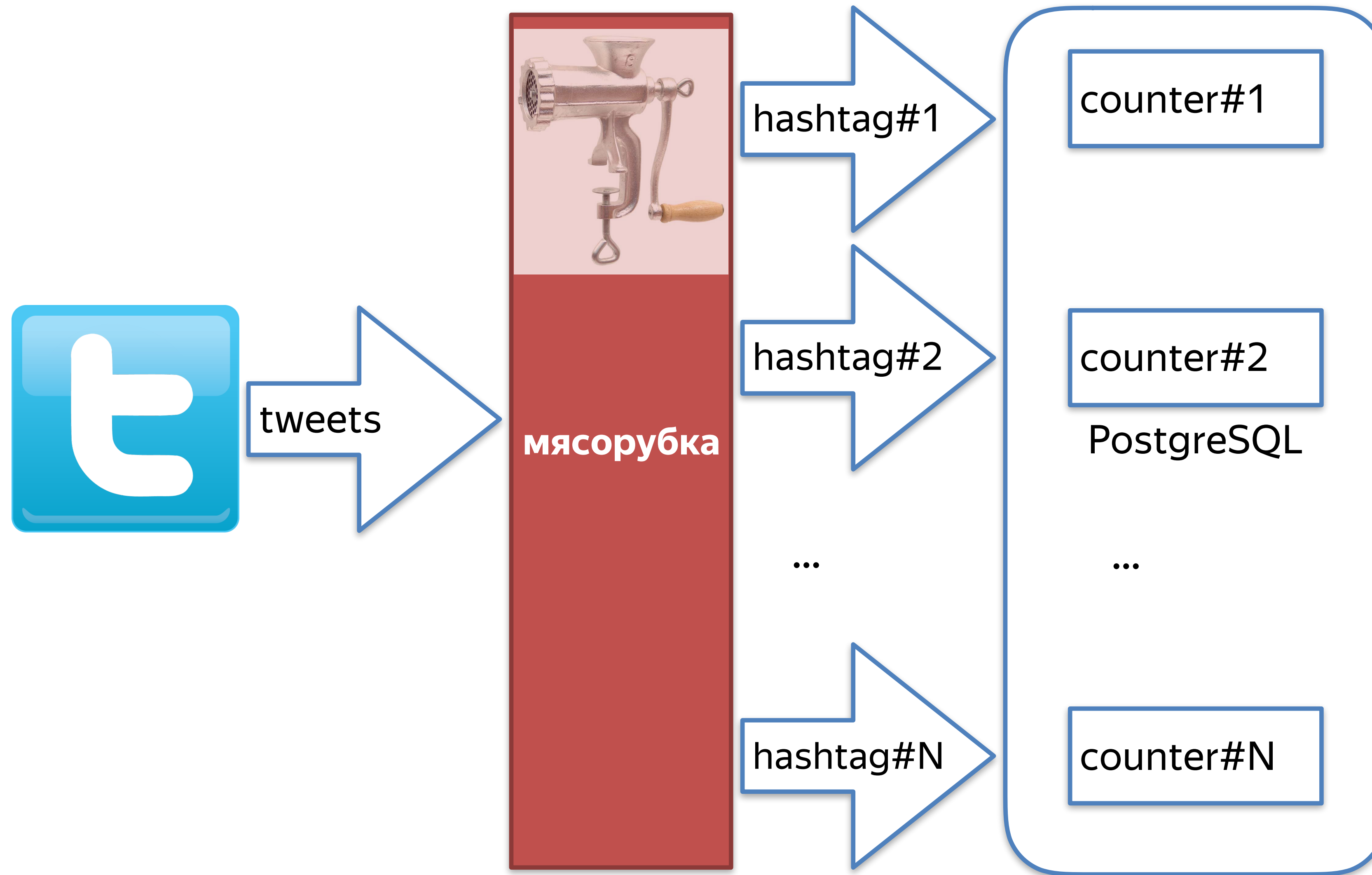
Инкрементальное решение



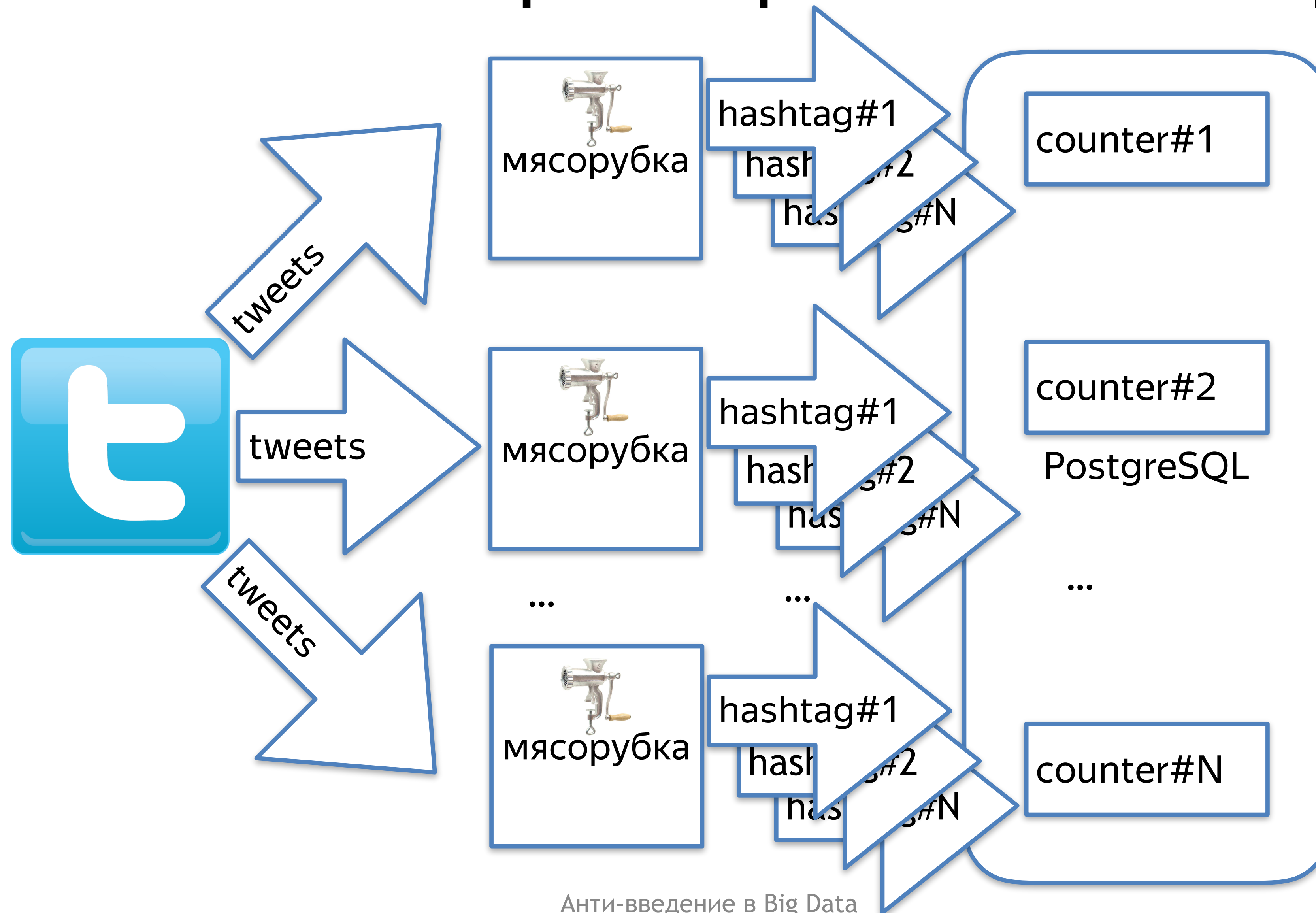
Инкрементальное решение



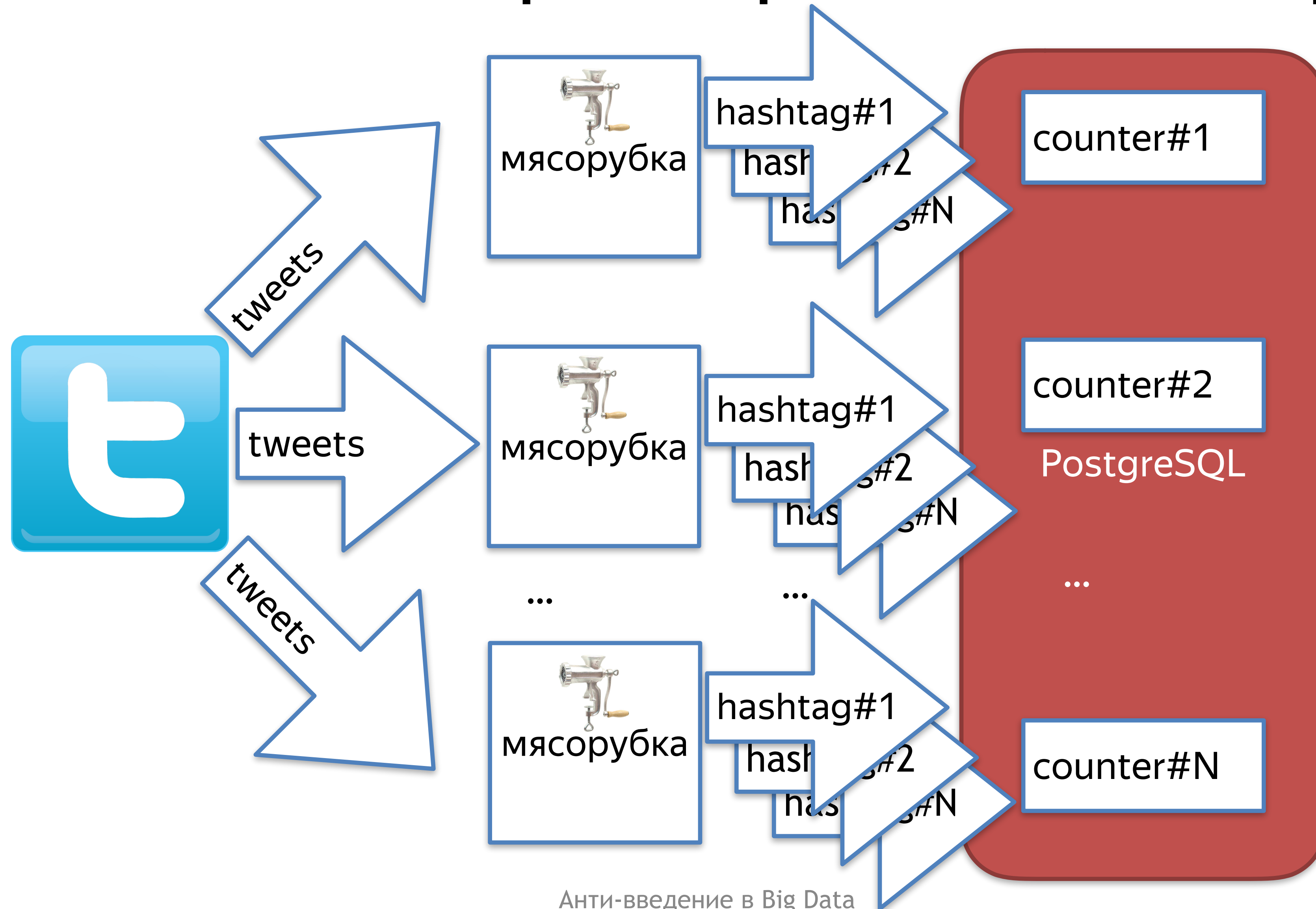
Инкрементальное решение



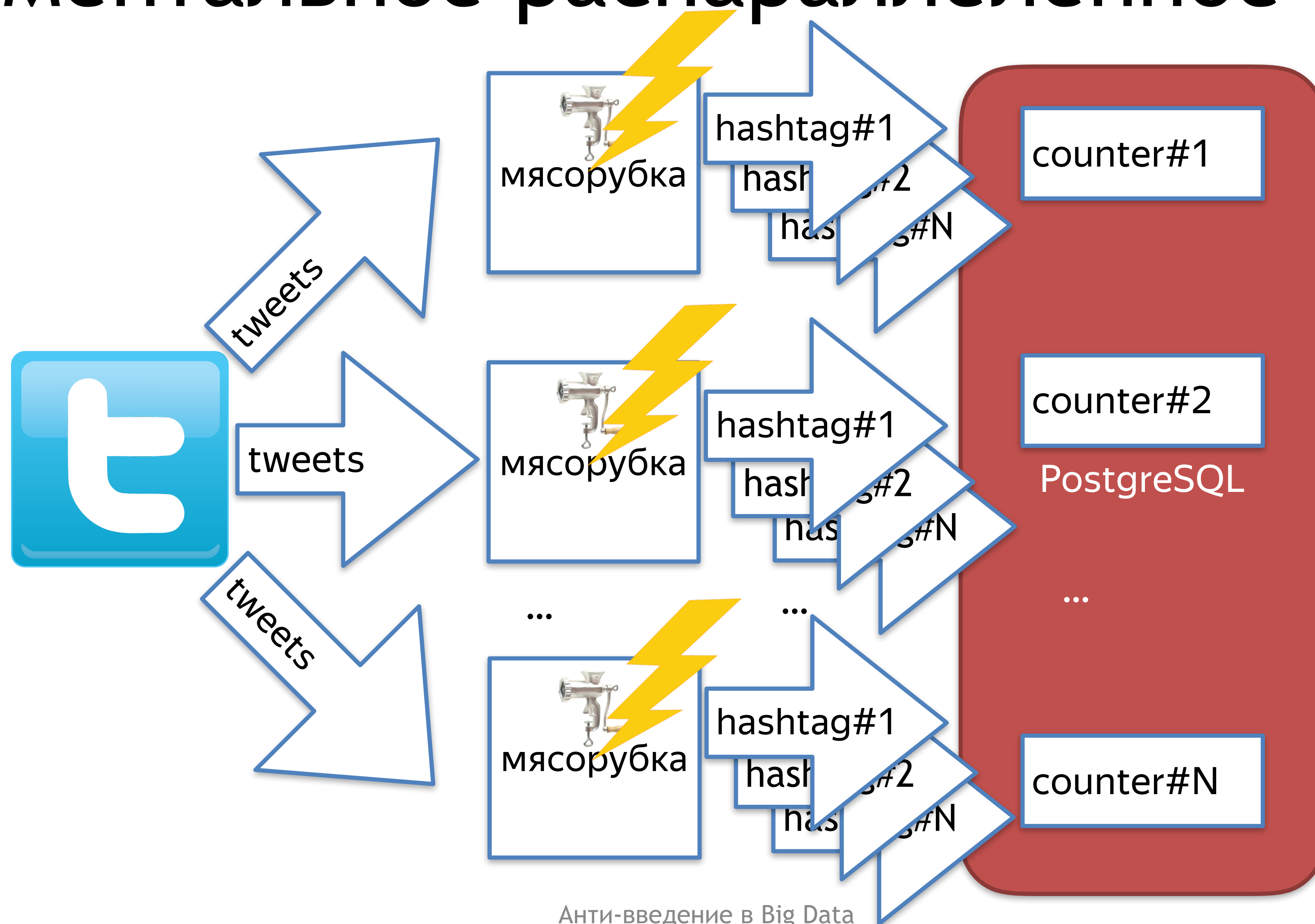
Инкрементальное распараллеленное решение



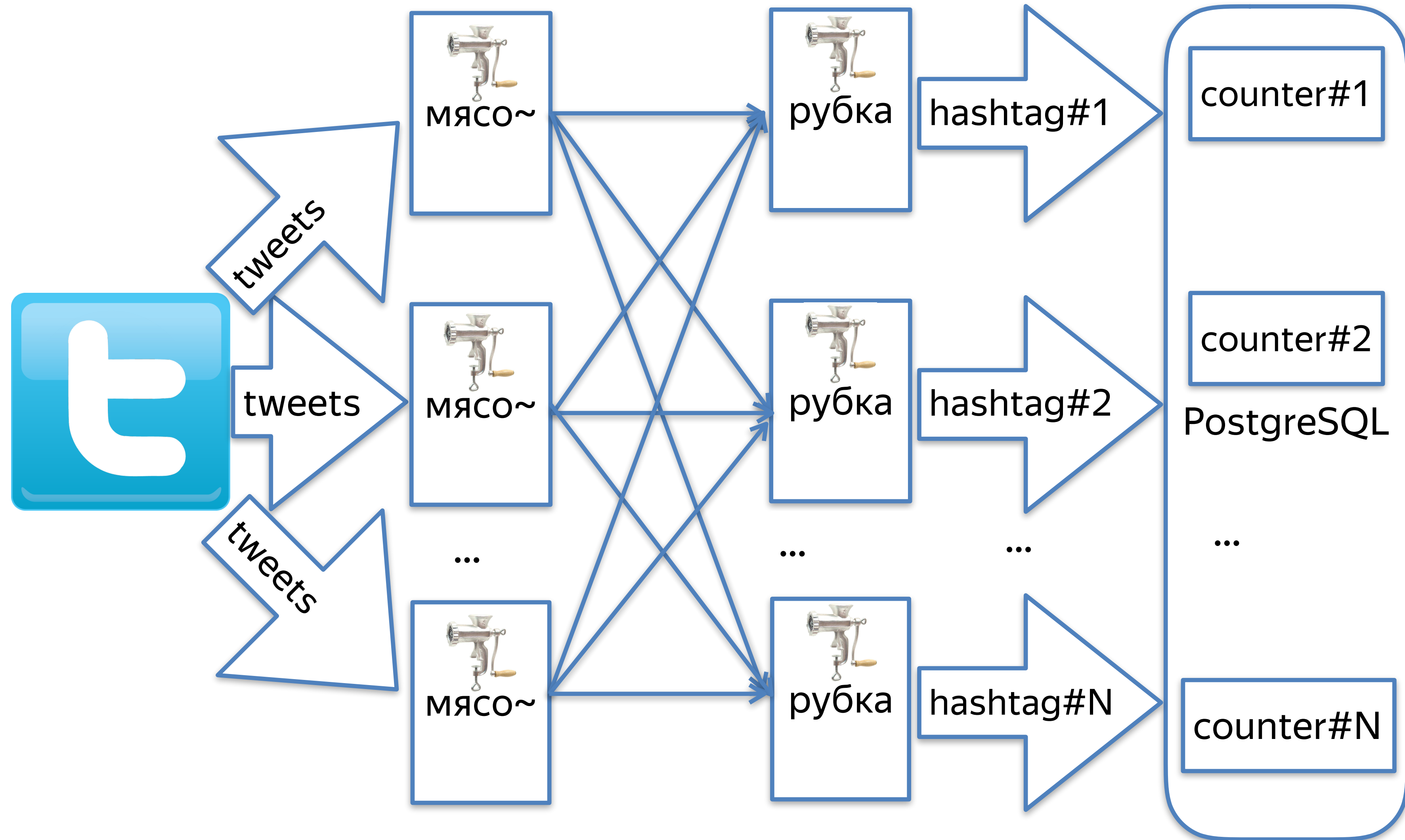
Инкрементальное распараллеленное решение



Инкрементальное распараллеленное решение



Инкрементальное распараллеленное решение без блокировок



Но есть две проблемы...



1. Решение плохо адаптируется под смежные задачи

1. Решение плохо адаптируется
под смежные задачи

> Регионы

1. Решение плохо адаптируется
под смежные задачи

> Регионы

> Временные интервалы

1. Решение плохо адаптируется под смежные задачи

- > Регионы
- > Временные интервалы
- > Другие источники хэштегов: VK, FB, Instagram

| 2. Люди допускают ашибки

~~2. Люди допускают ашибки~~
2. Люди допускают ошибки

Мясорубка

```
List list = tweet.getHashtags();  
list.forEach(hashtag -> {  
    long counter = getCurrentCounter(hashtag);  
    counter++;  
    setCurrentCounter(hashtag, counter);  
});
```

Мисопубка

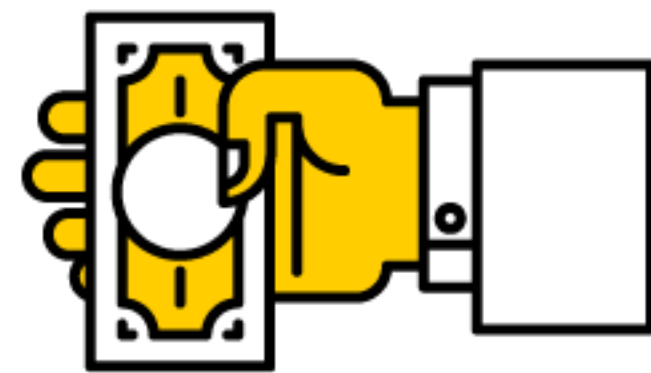
```
List list = tweet.getHashtags();  
list.forEach(hashtag -> {  
    long counter = getCurrentCounter(hashtag);  
    counter--;  
    setCurrentCounter(hashtag, counter);  
});
```


умные-решения.рф

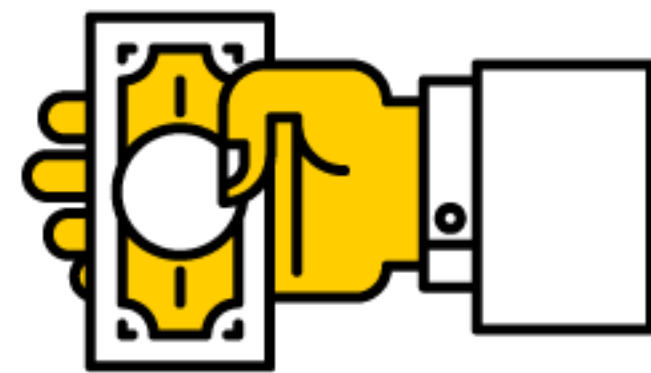
умные-решения.рф



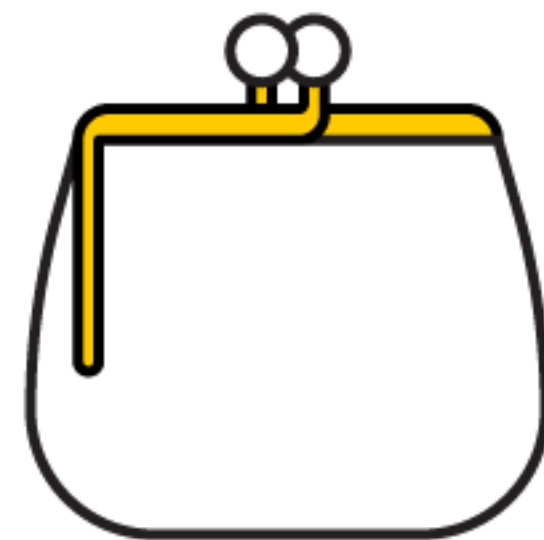
умные-решения.рф



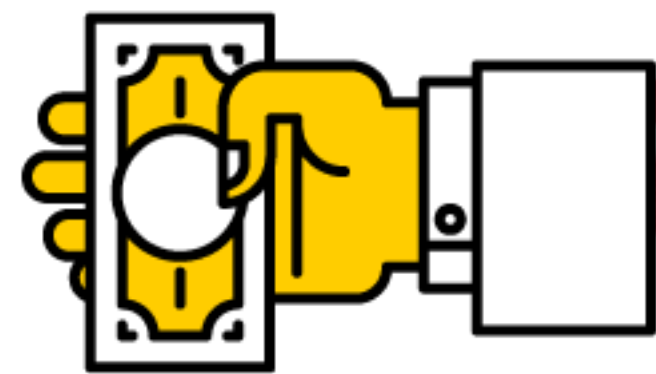
умные-решения.рф



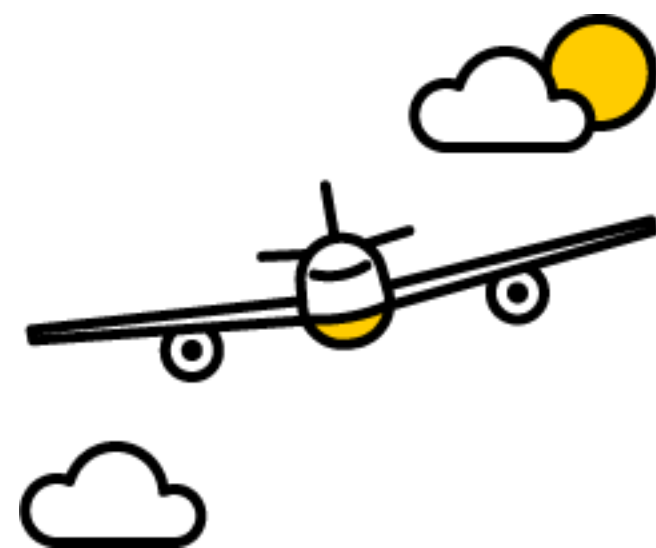
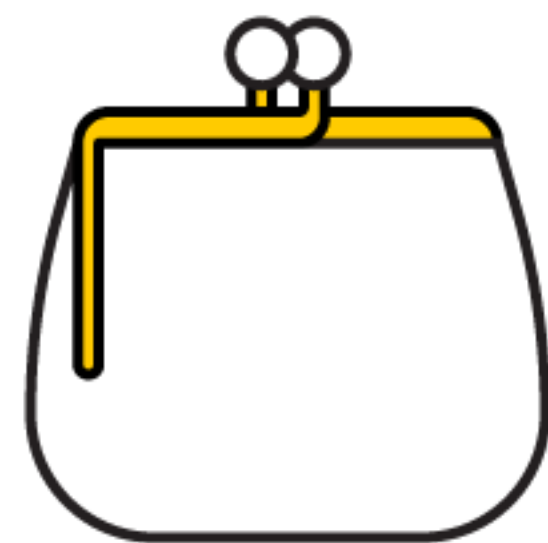
BANKRUPT



умные-решения.рф



BANKRUPT



более-умные-решения.рф

более-умные-решения.рф

1. Отвечает на вопрос какие сейчас* самые популярные хештеги.

более-умные-решения.рф

1. Отвечает на вопрос какие сейчас* самые популярные хештеги.
2. Хорошо адаптируется под смежные задачи.

более-умные-решения.рф

1. Отвечает на вопрос какие сейчас* самые популярные хештеги.
2. Хорошо адаптируется под смежные задачи.
3. Устойчиво к человеческим ошибкам.

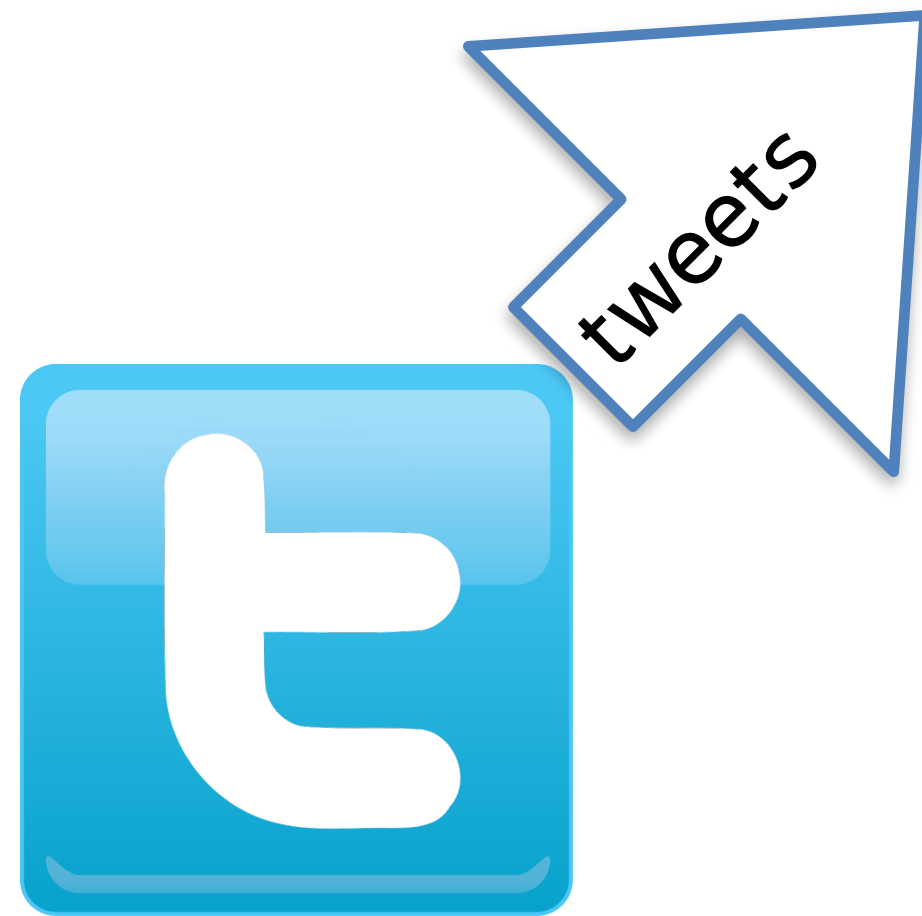
Более умное решение

Более умное решение

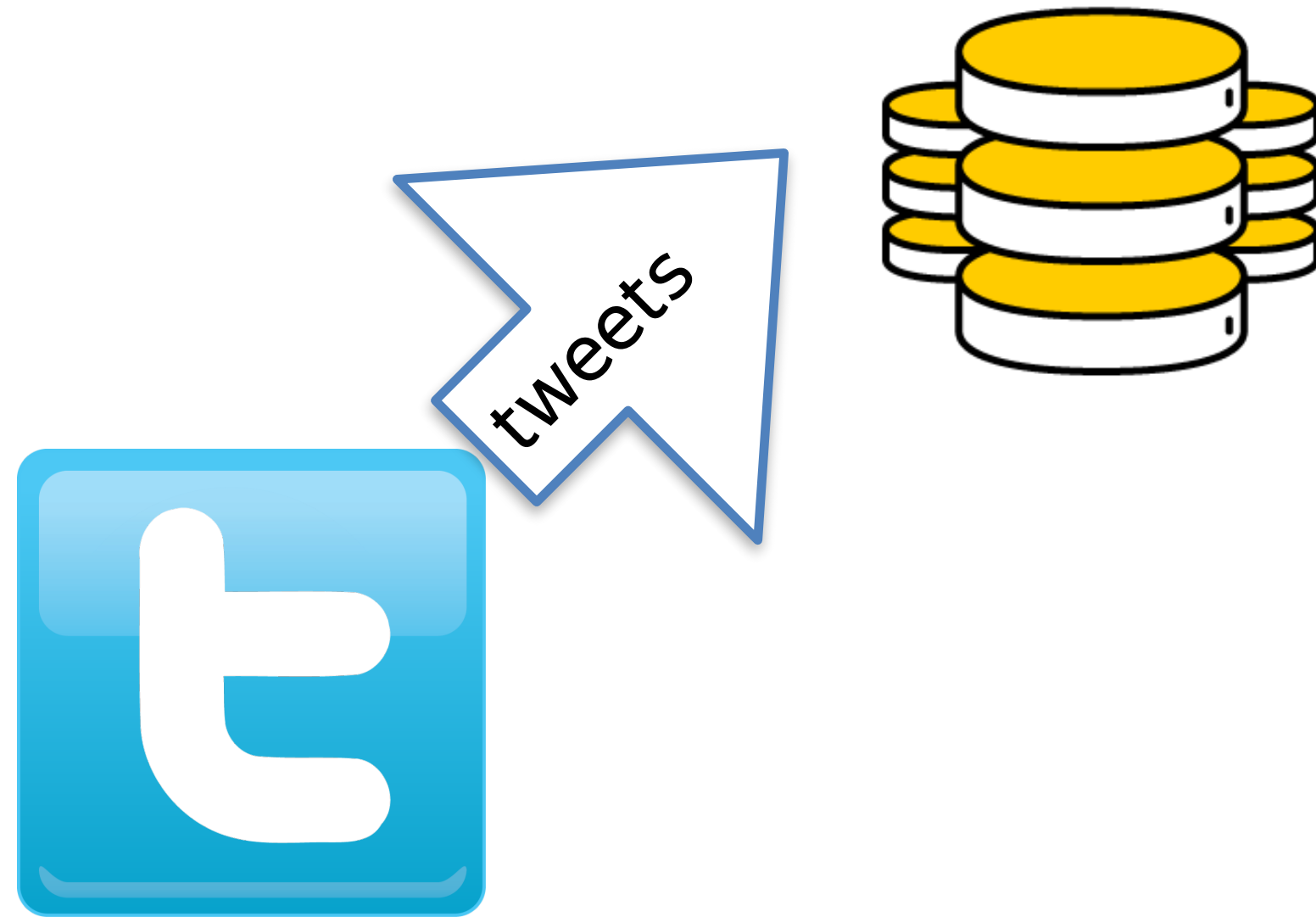
Более умное решение



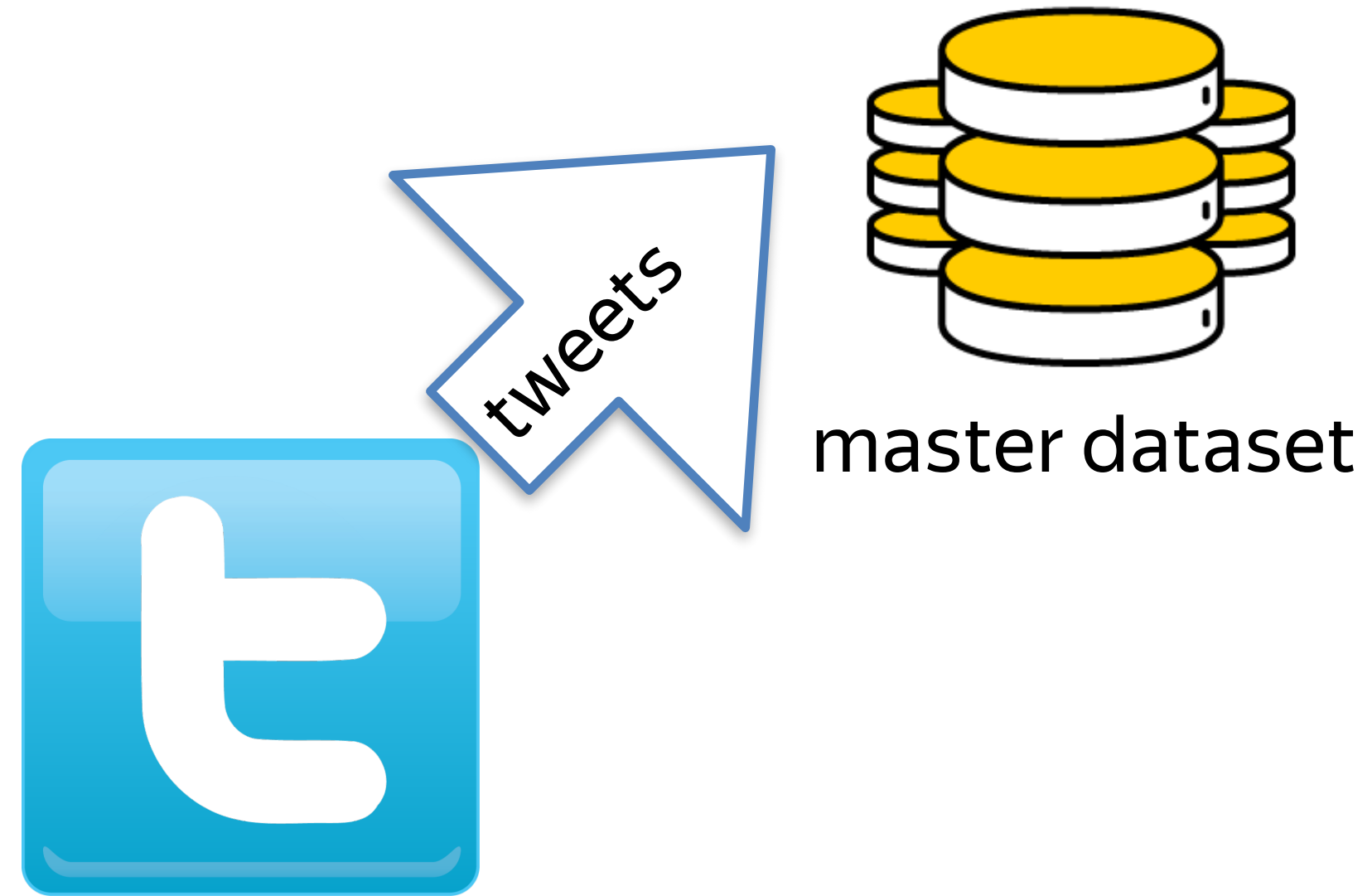
Более умное решение



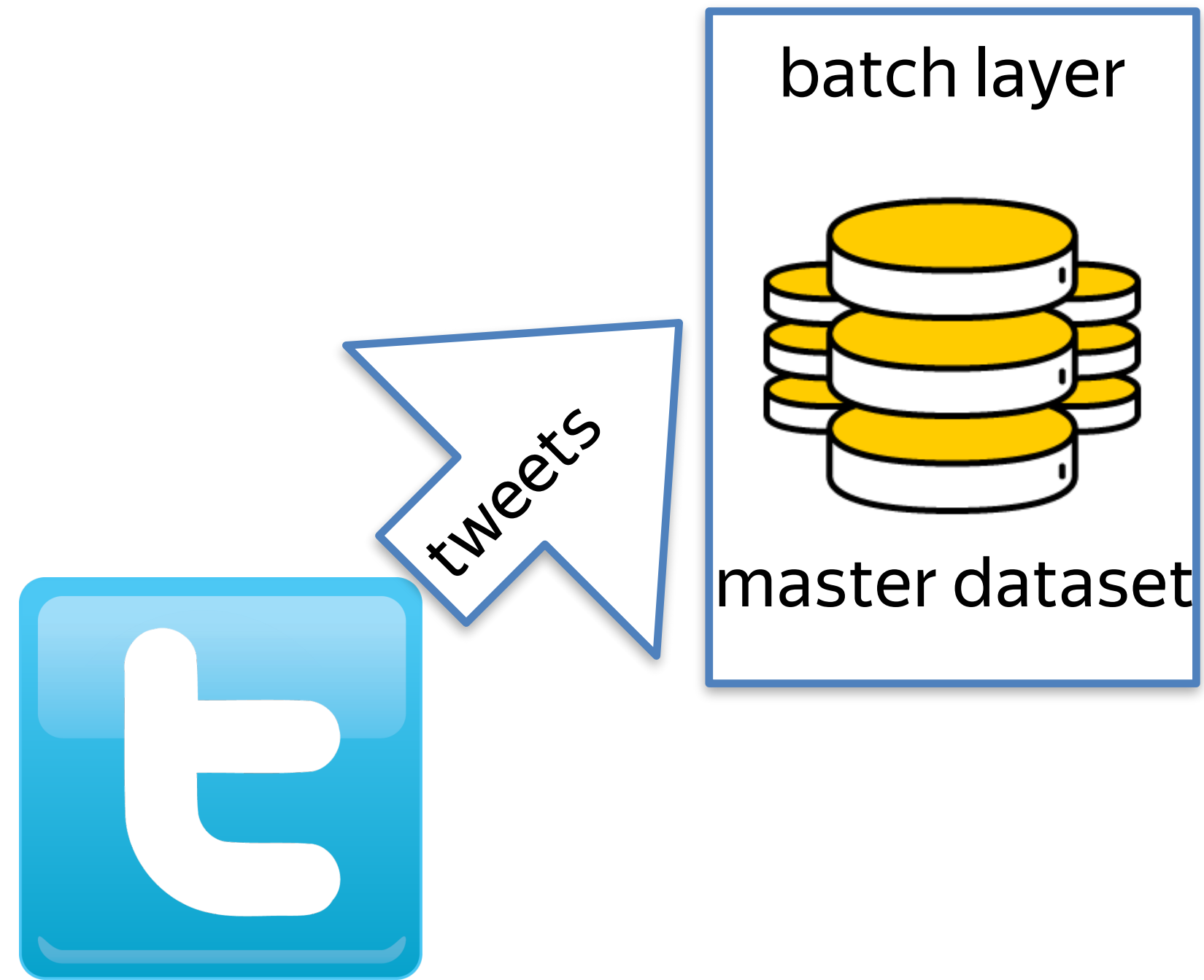
Более умное решение



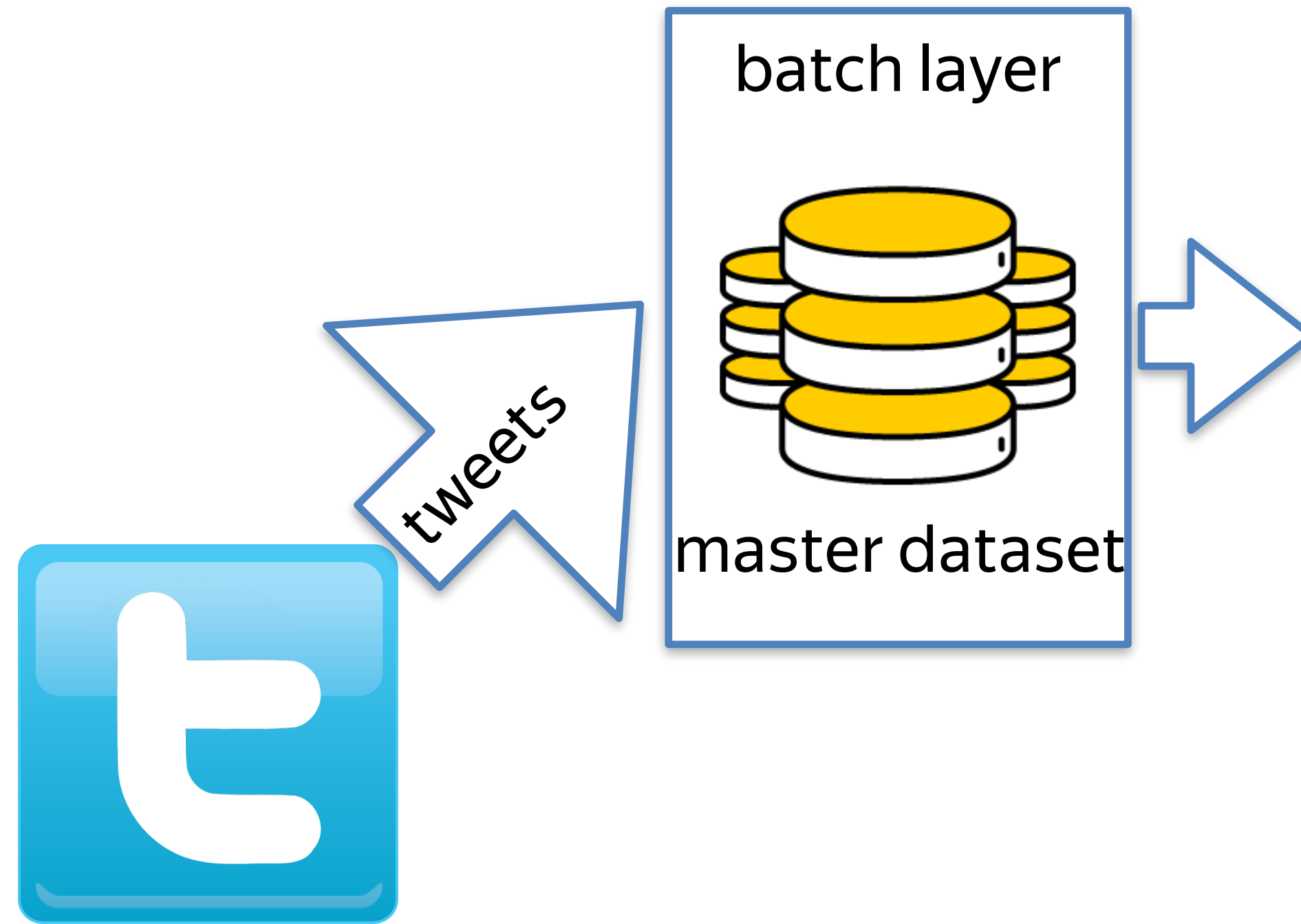
Более умное решение



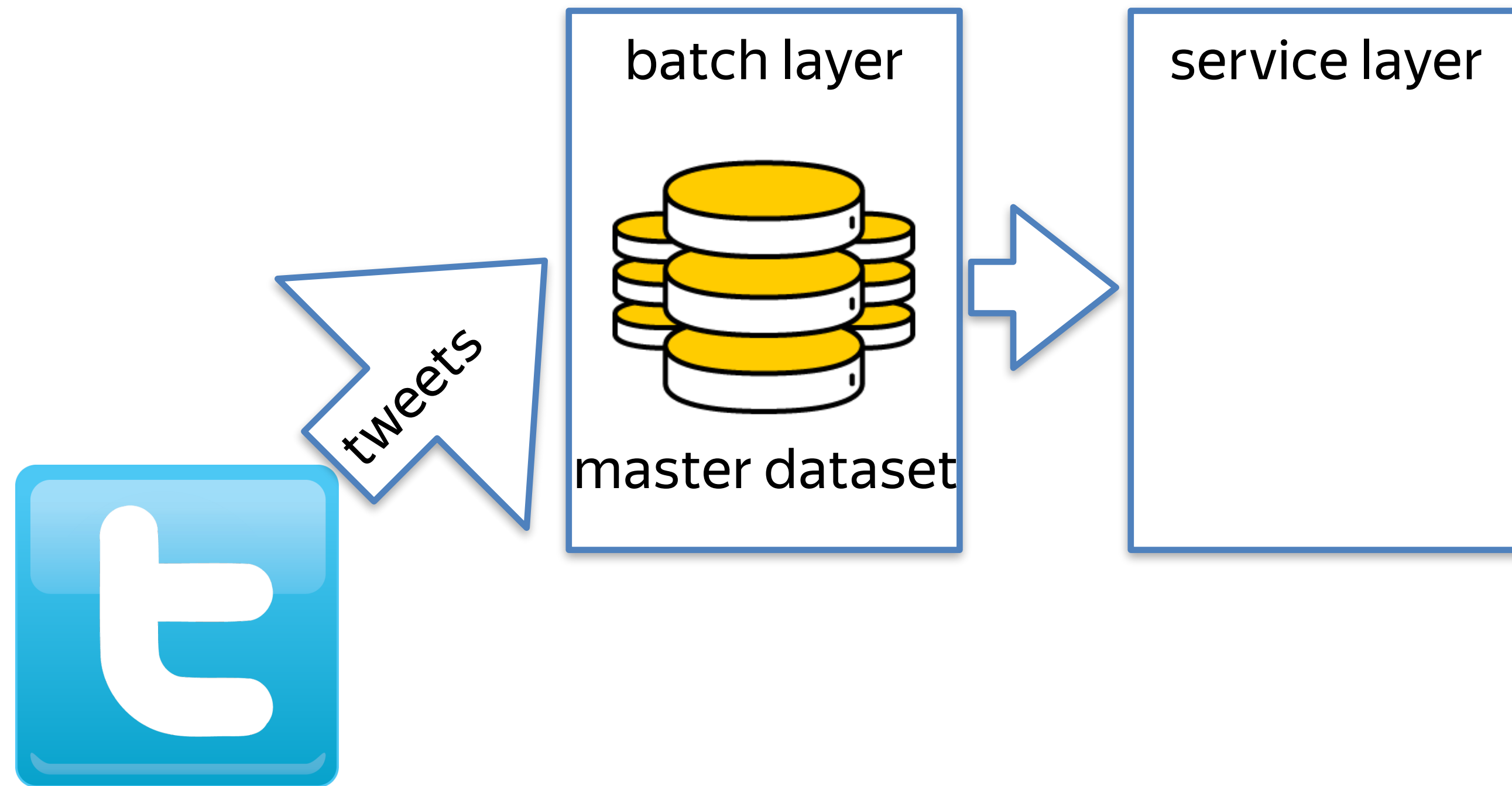
Более умное решение



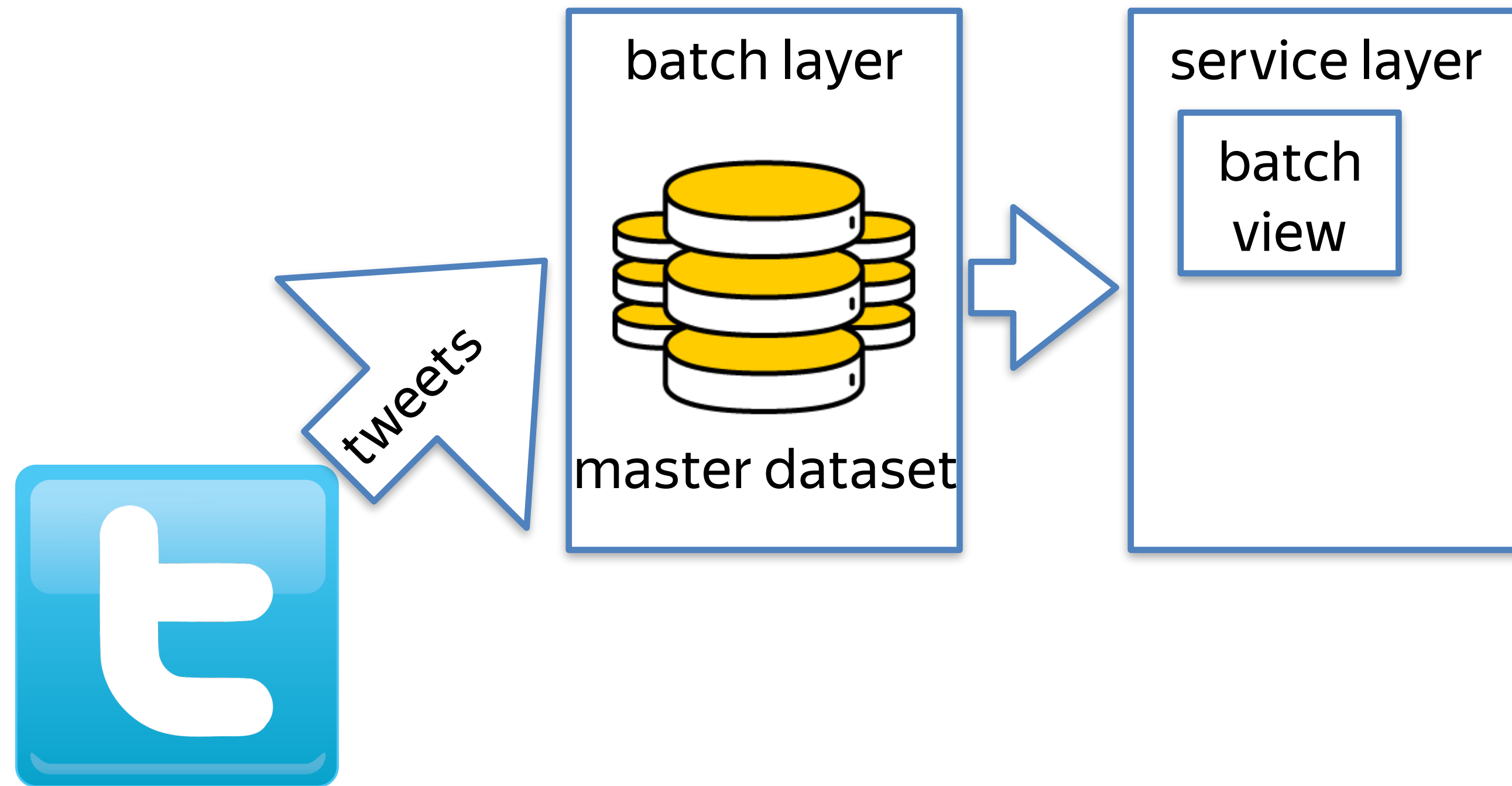
Более умное решение



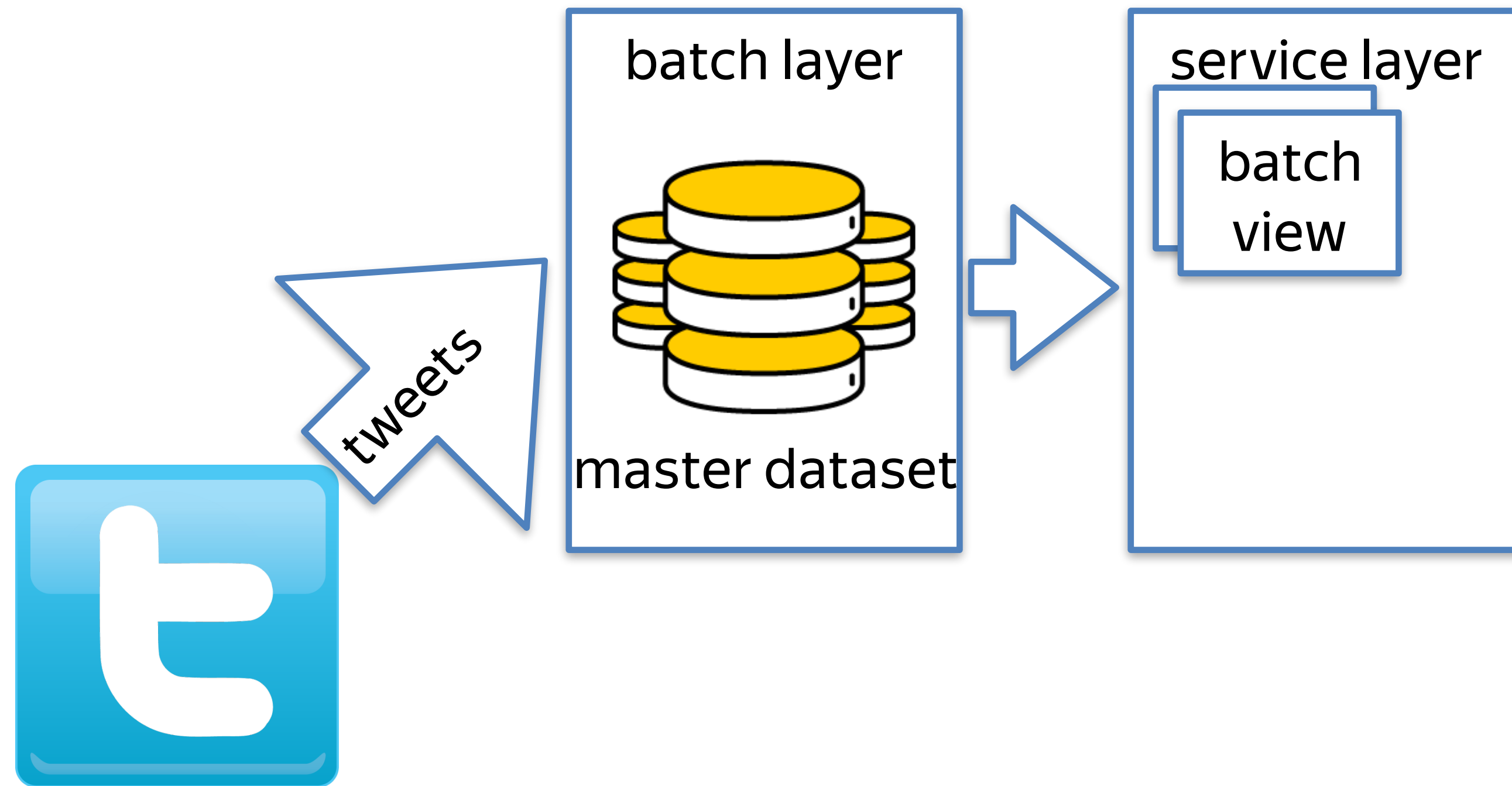
Более умное решение



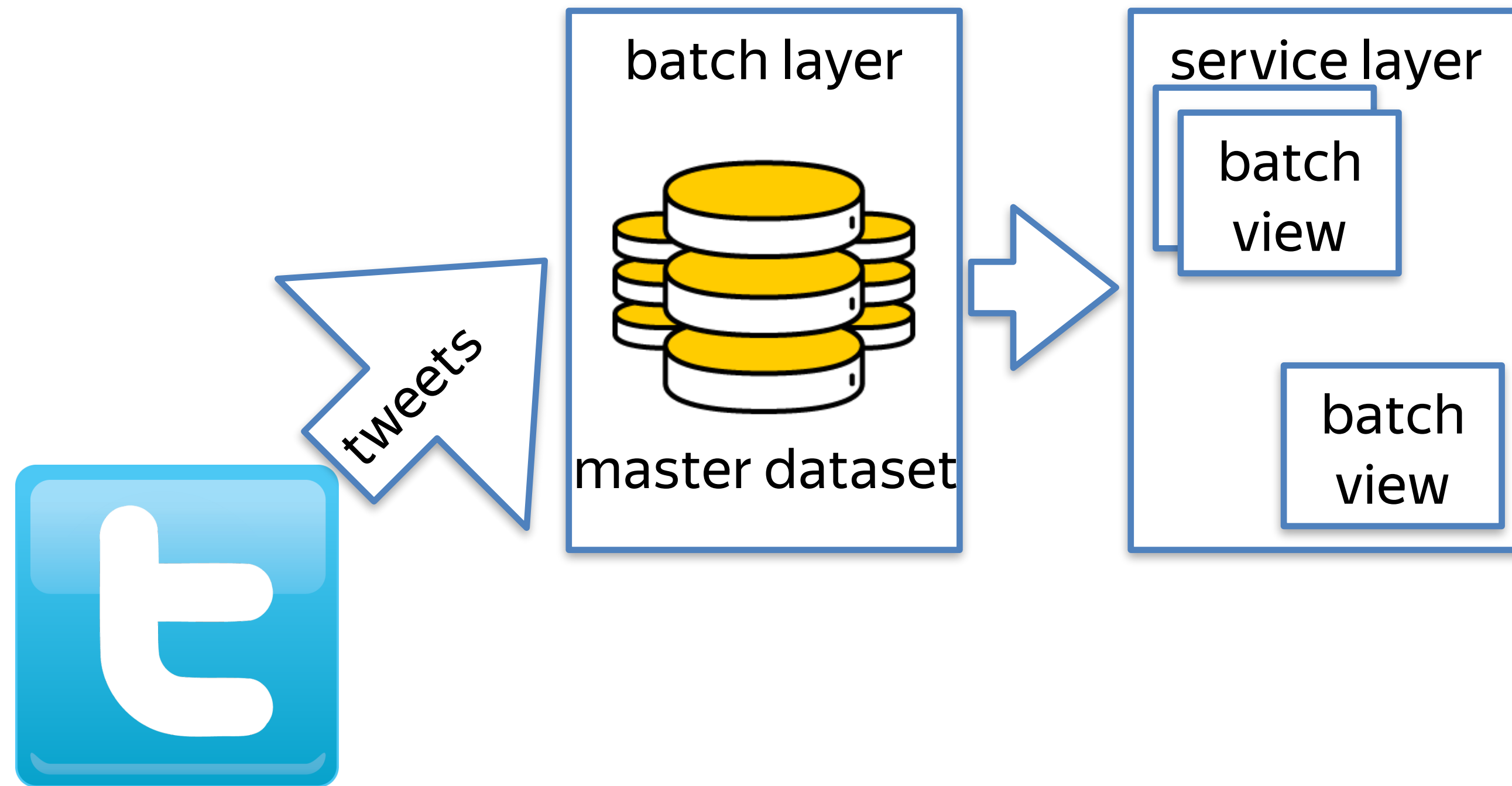
Более умное решение



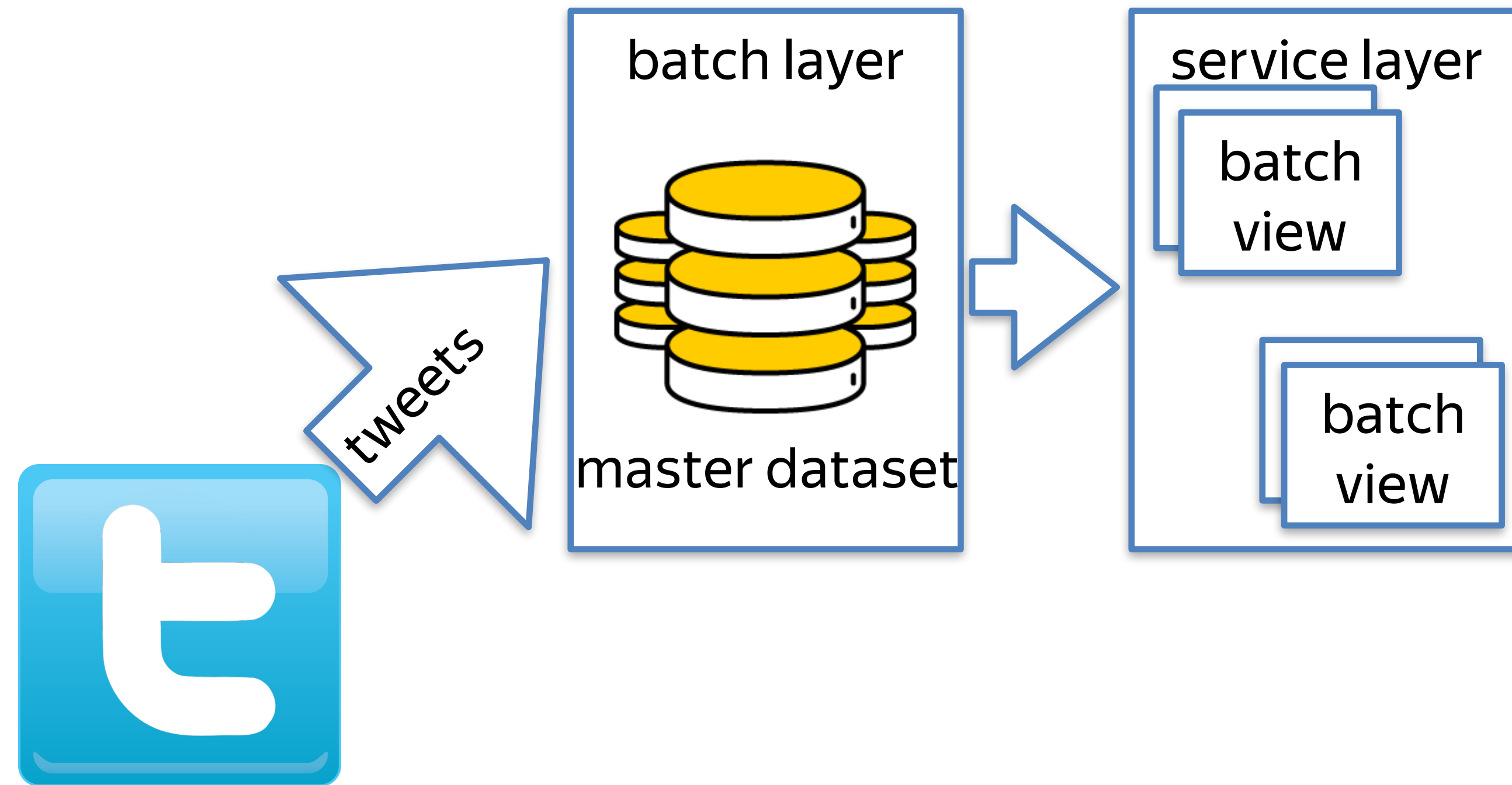
Более умное решение



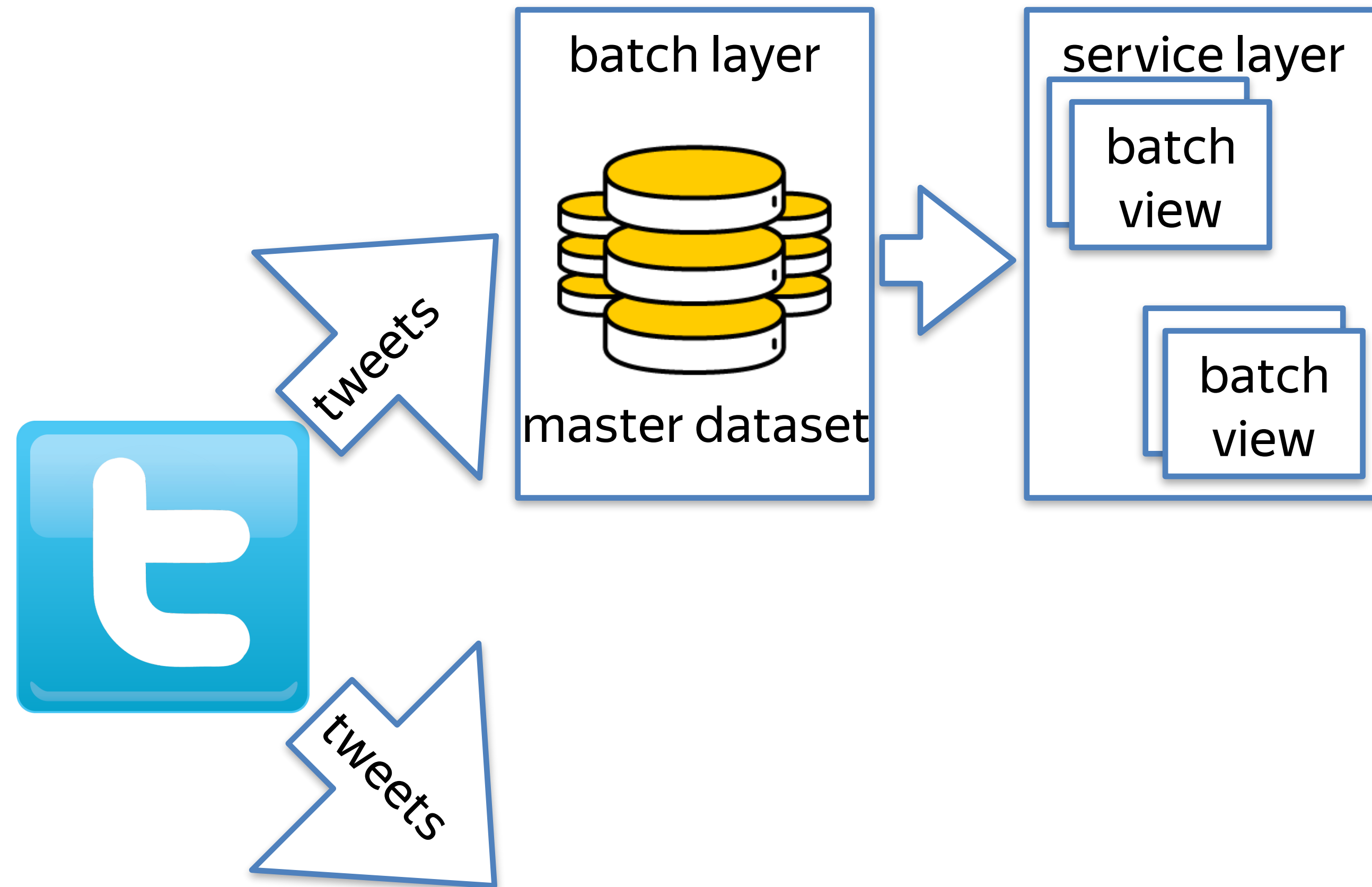
Более умное решение



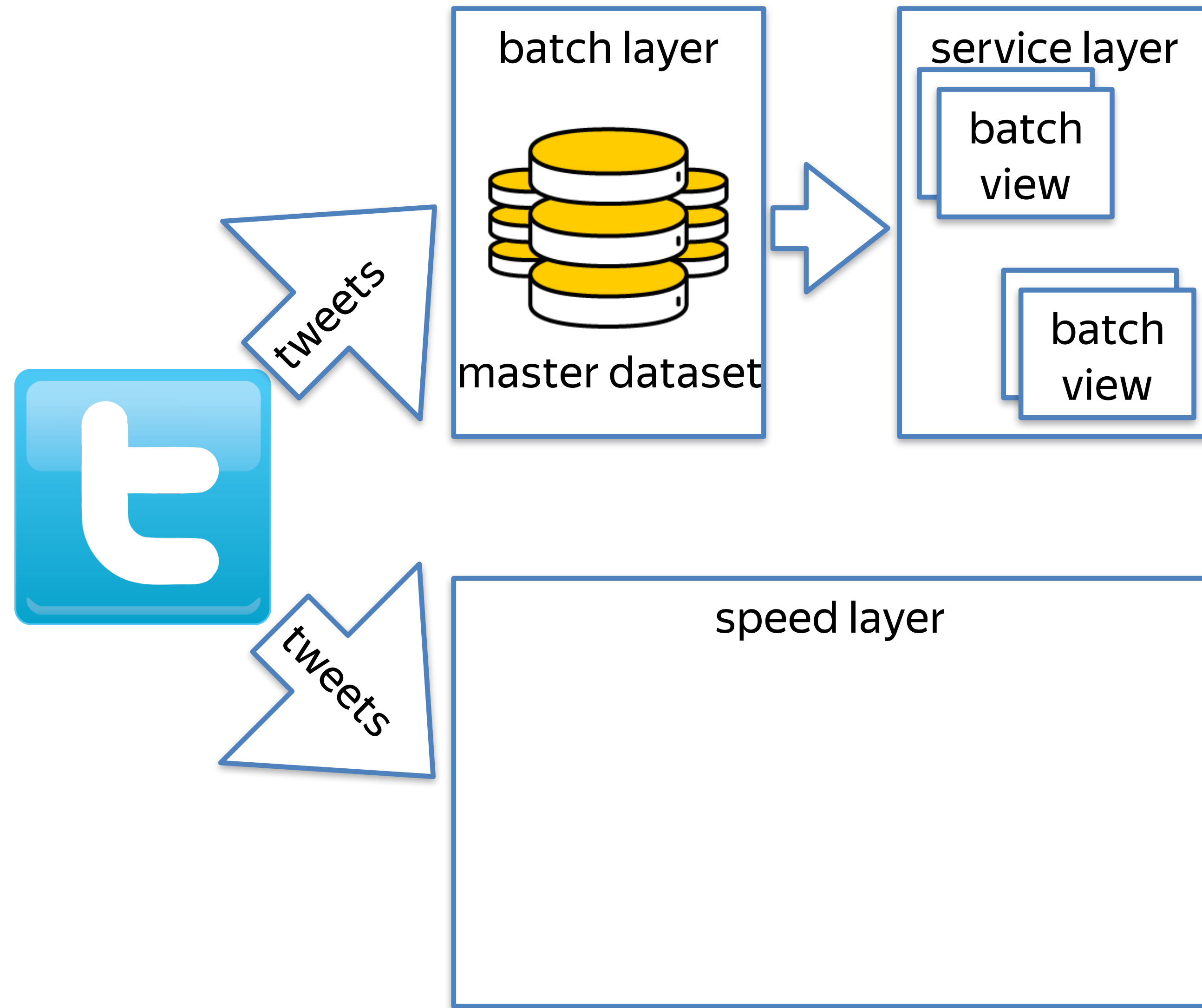
Более умное решение



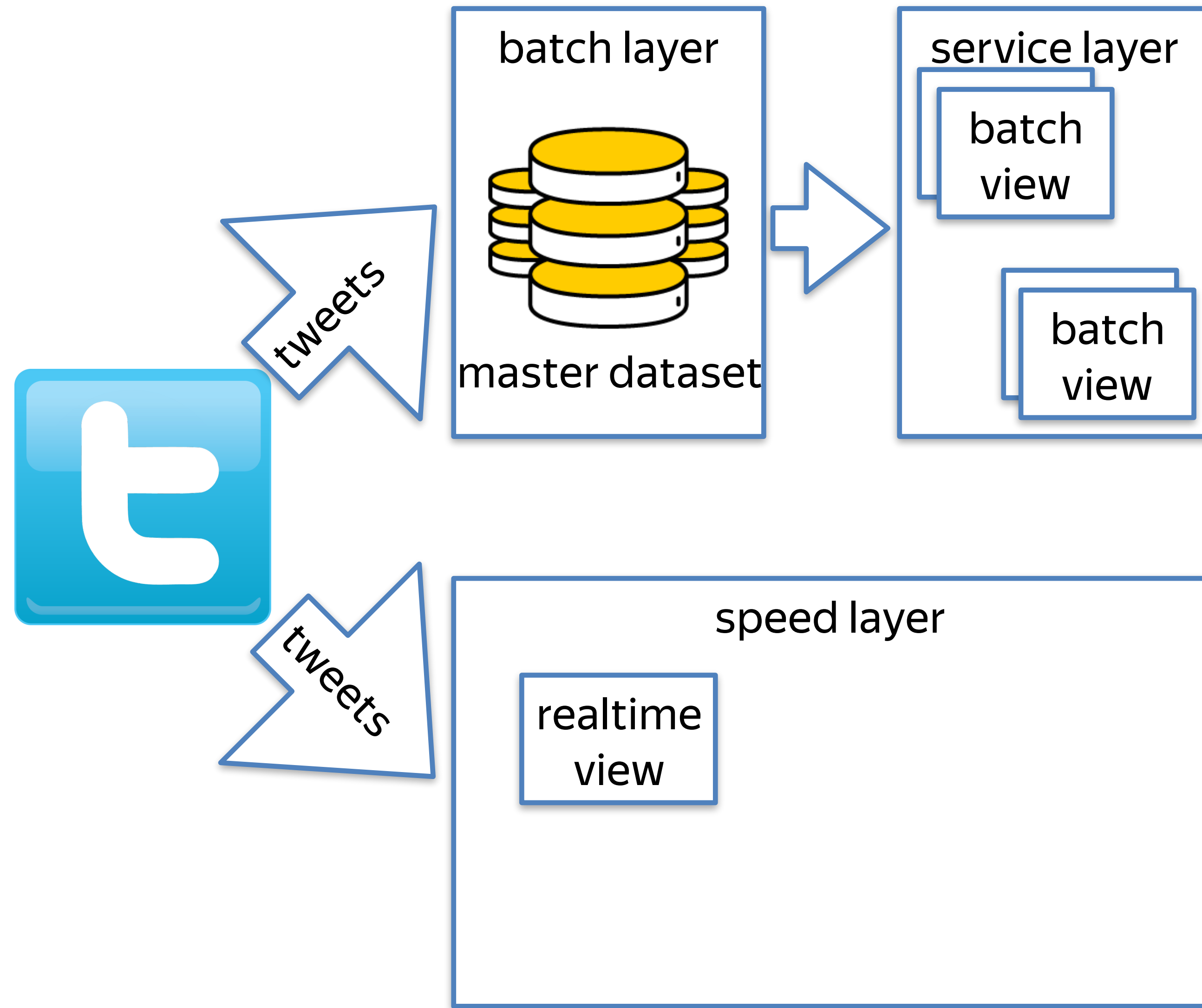
Более умное решение



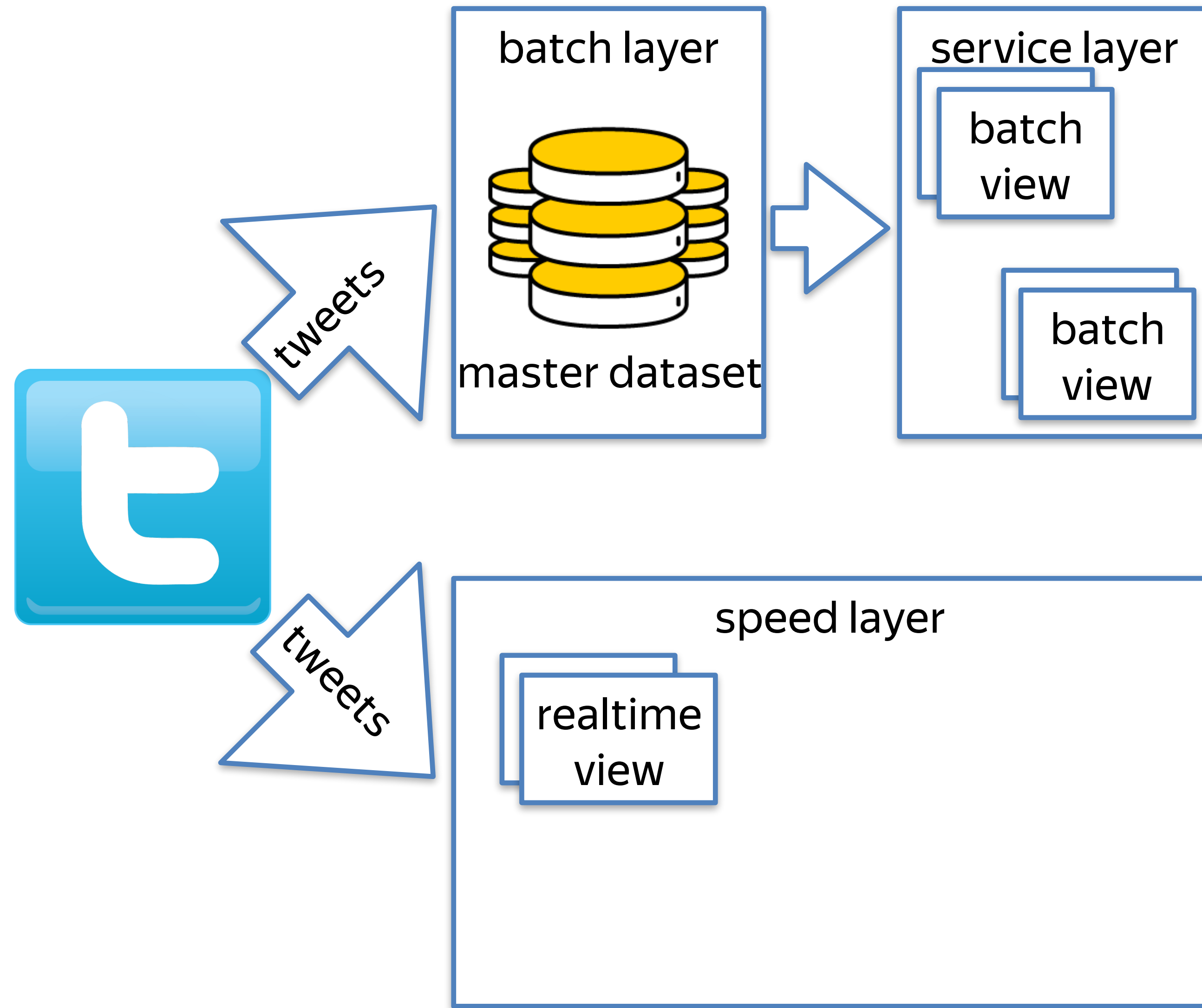
Более умное решение



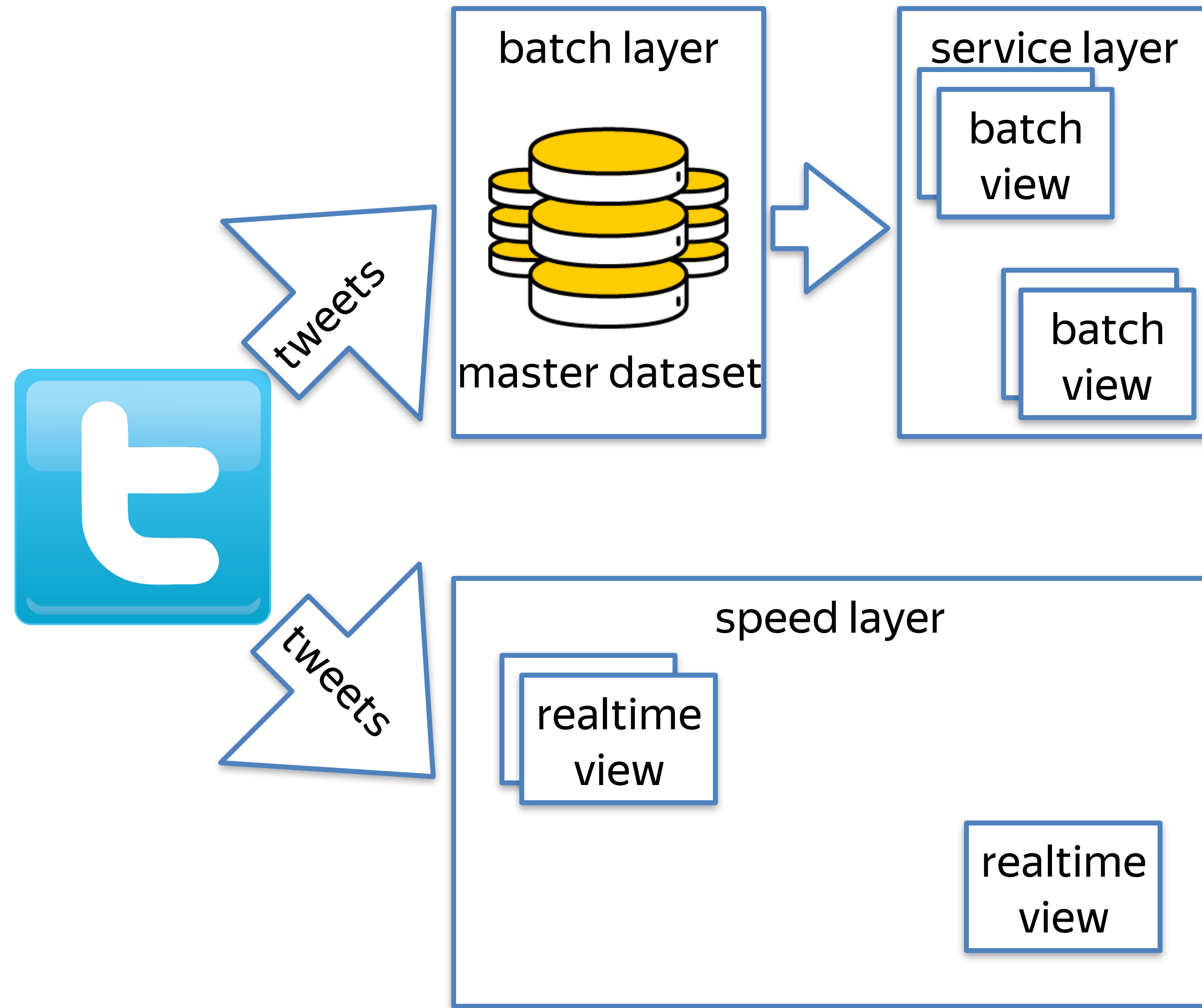
Более умное решение



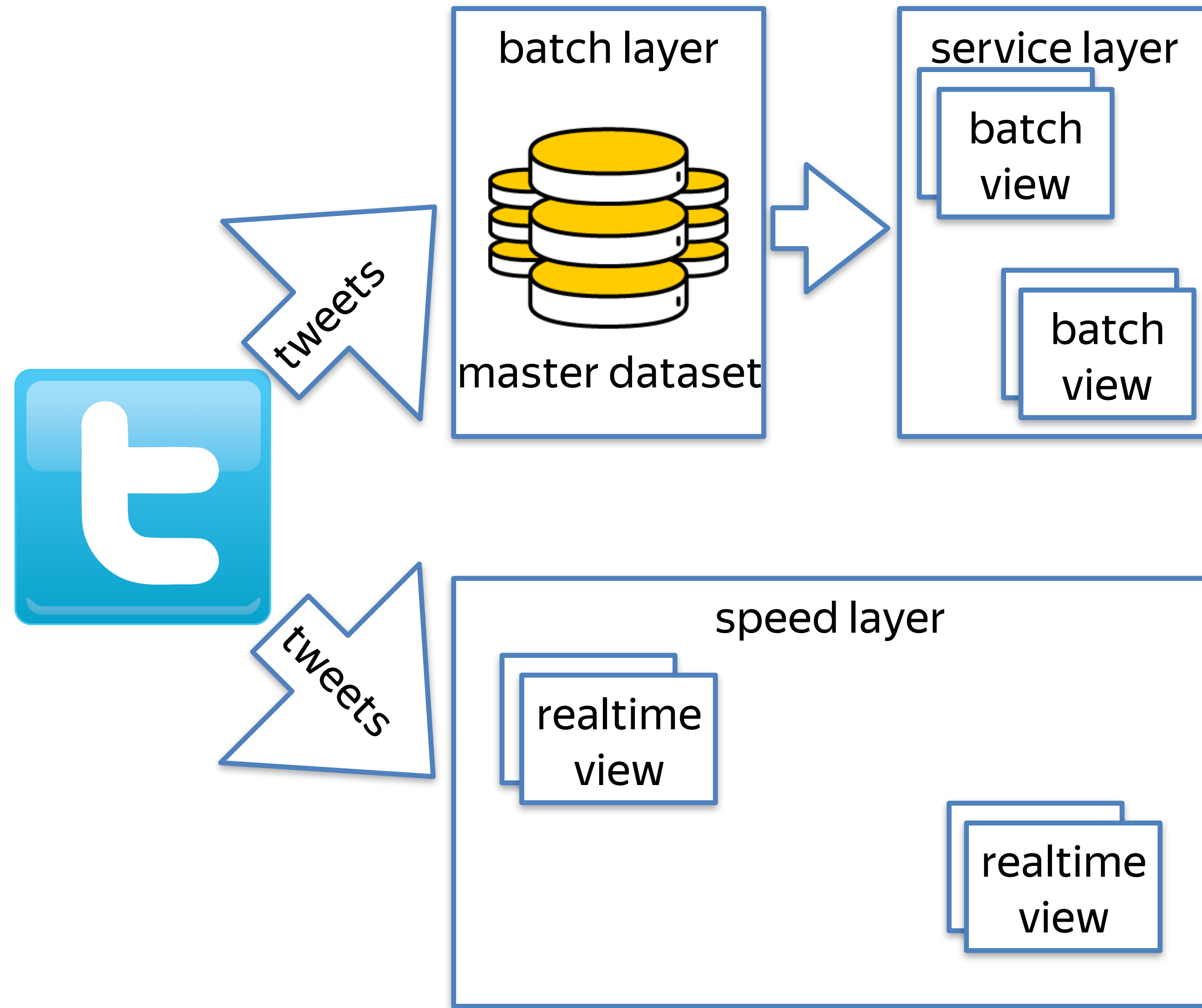
Более умное решение



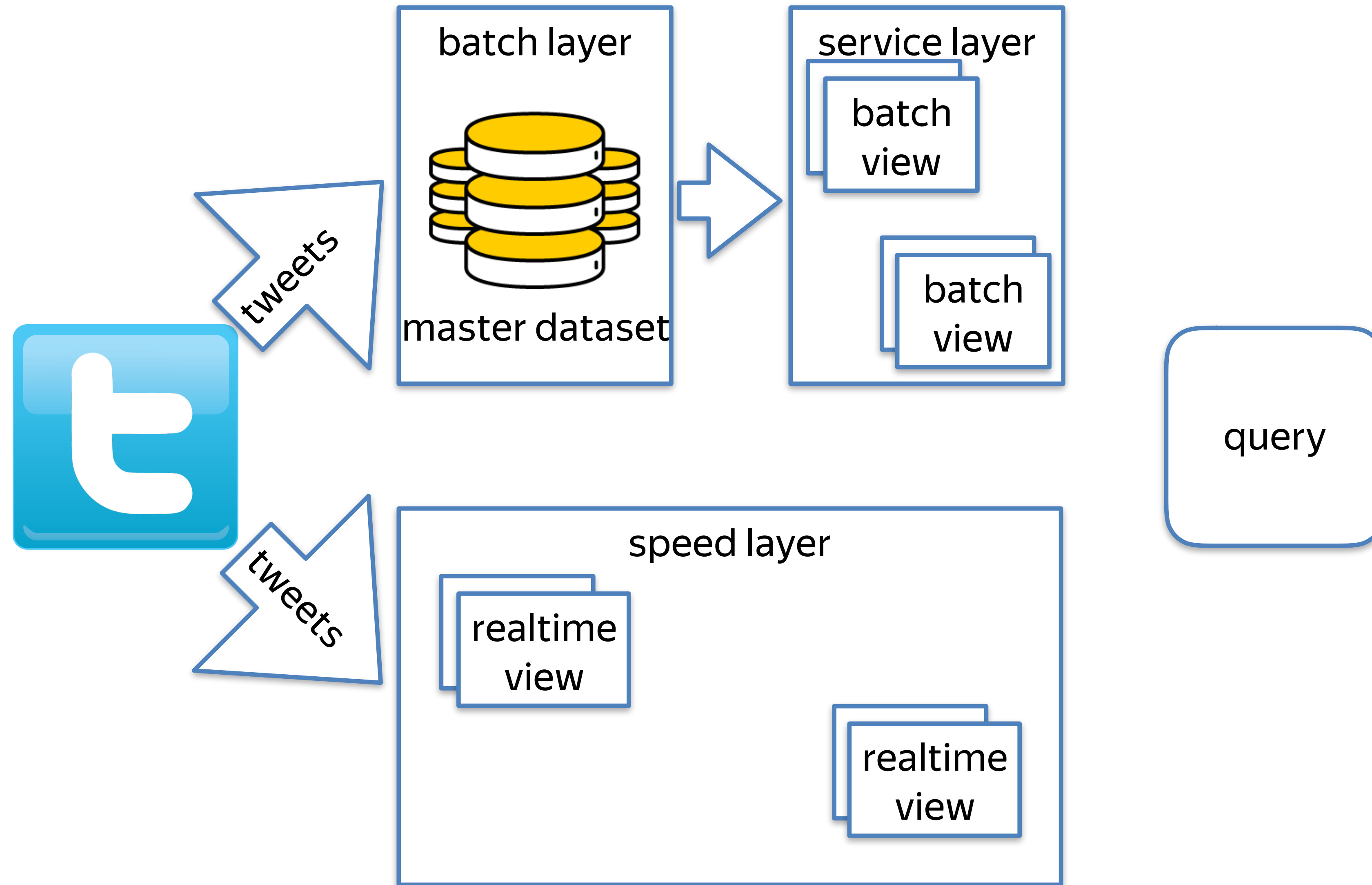
Более умное решение



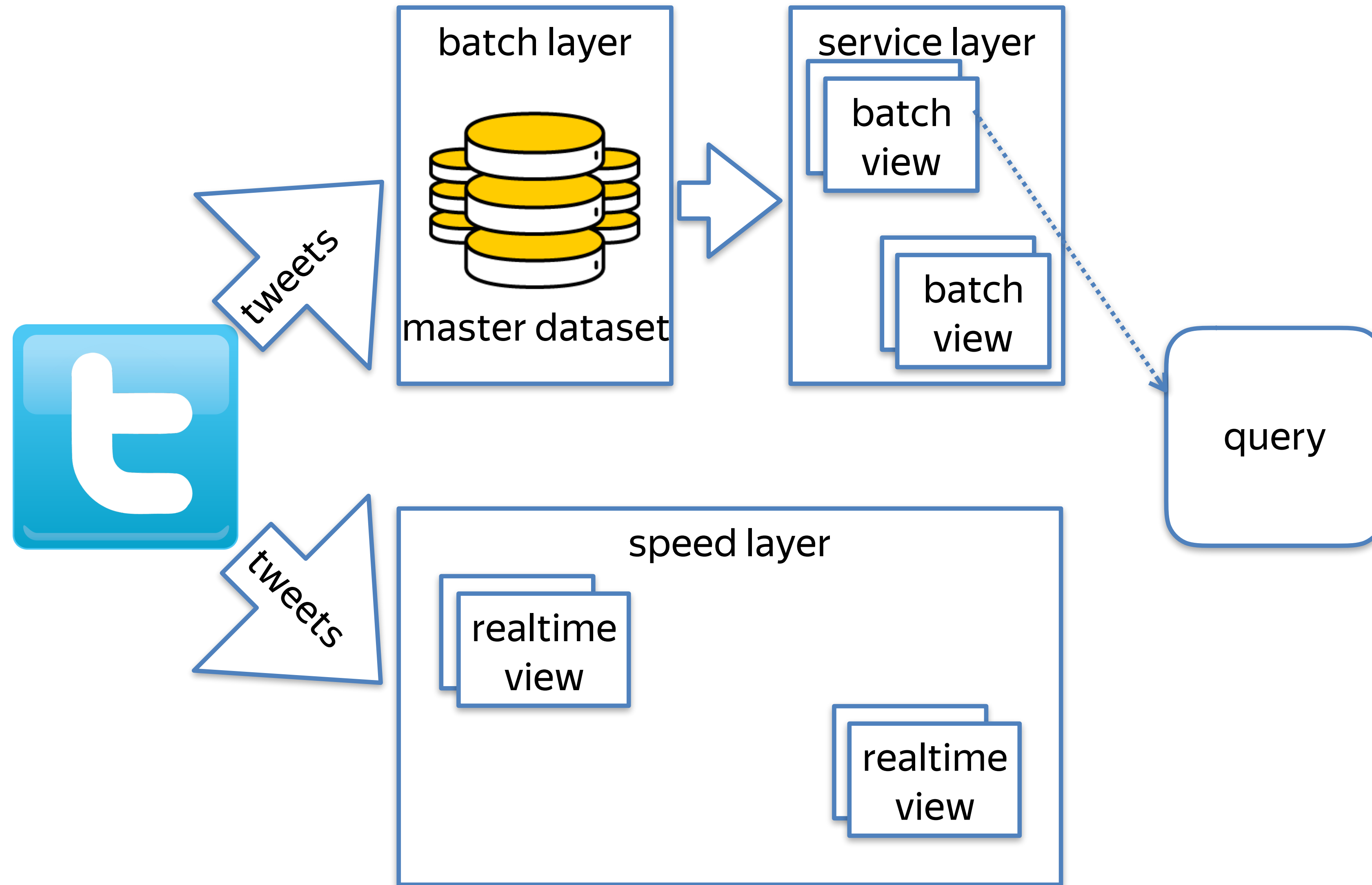
Более умное решение



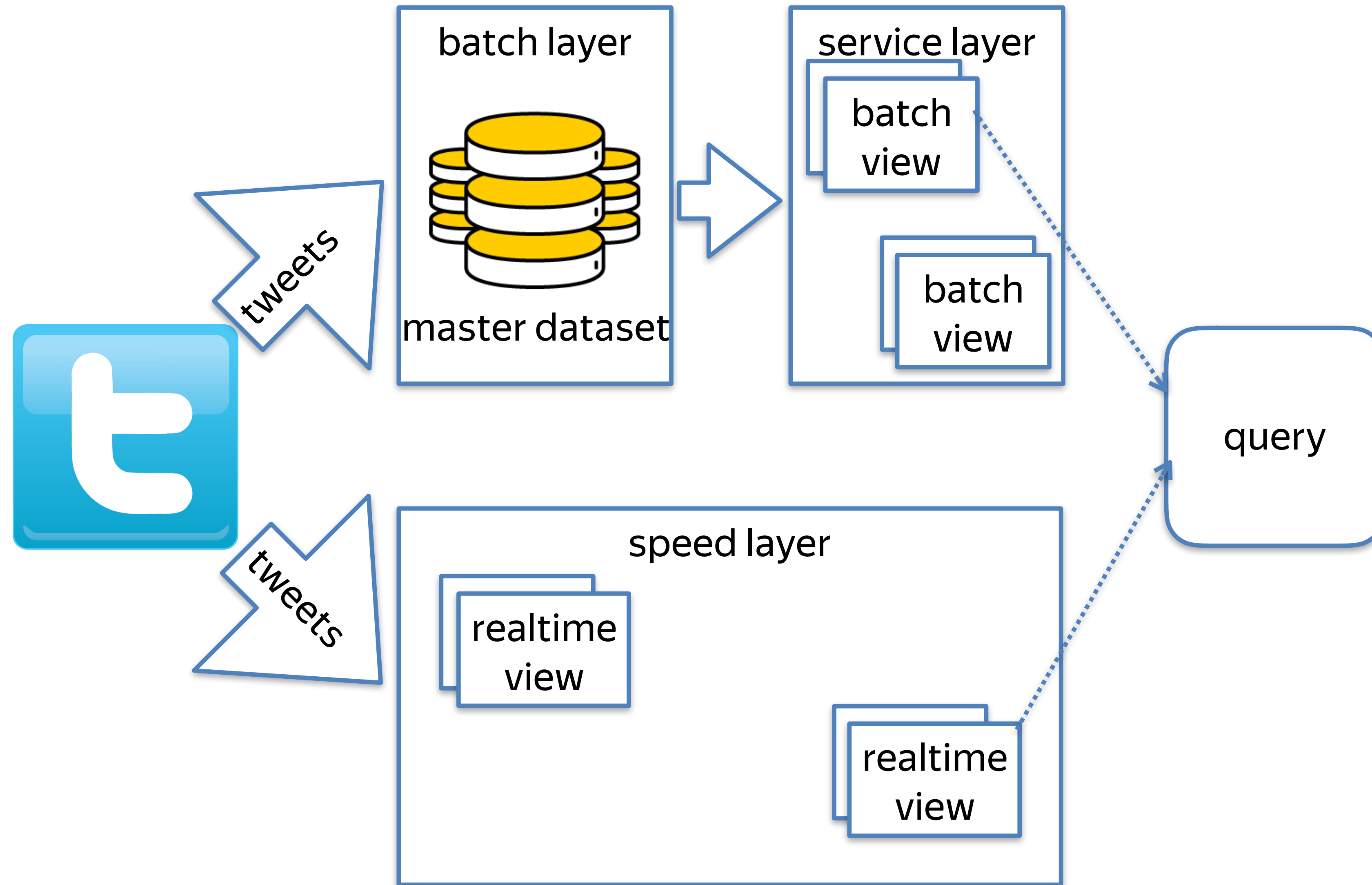
Более умное решение



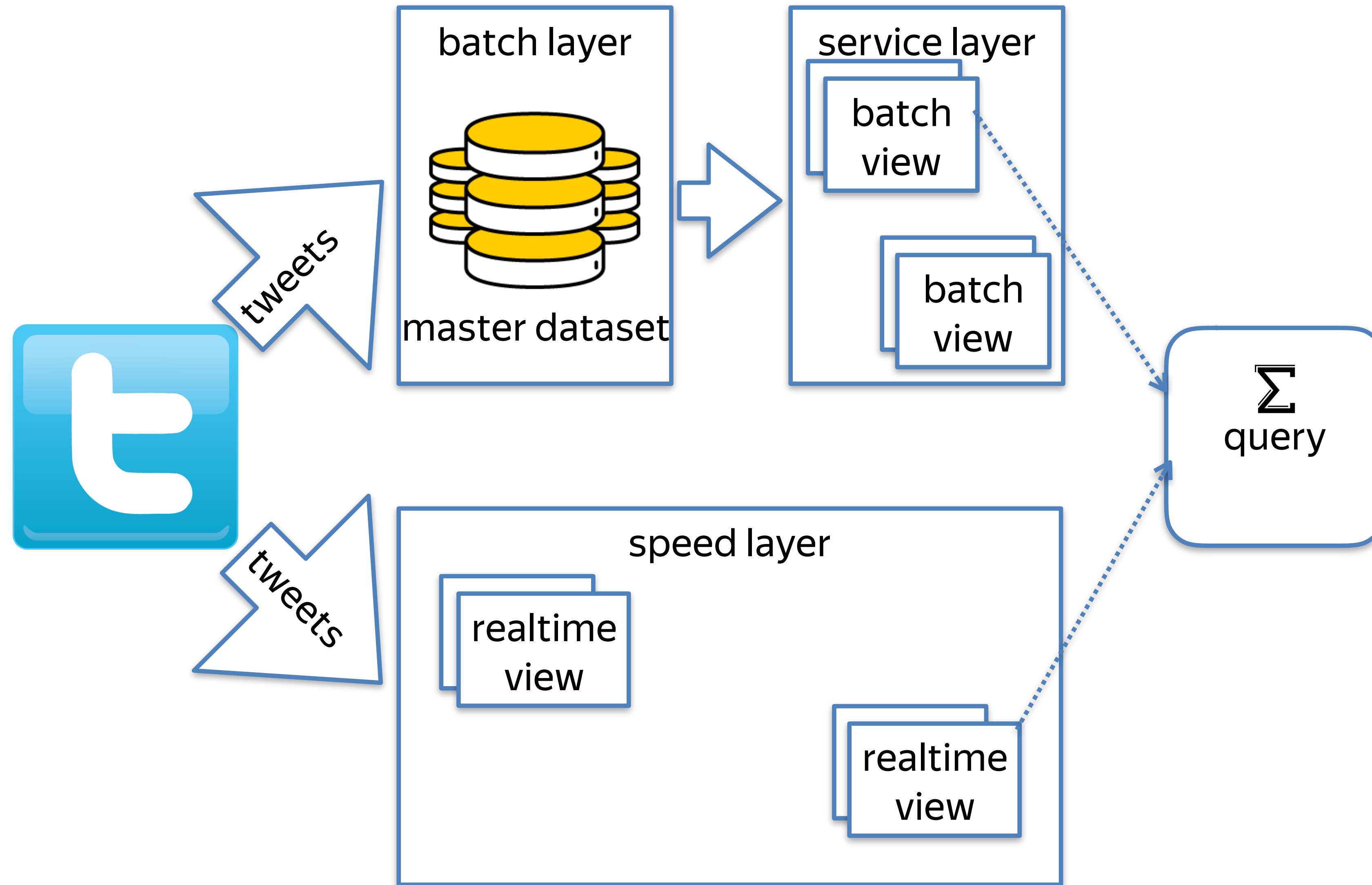
Более умное решение



Более умное решение



Более умное решение

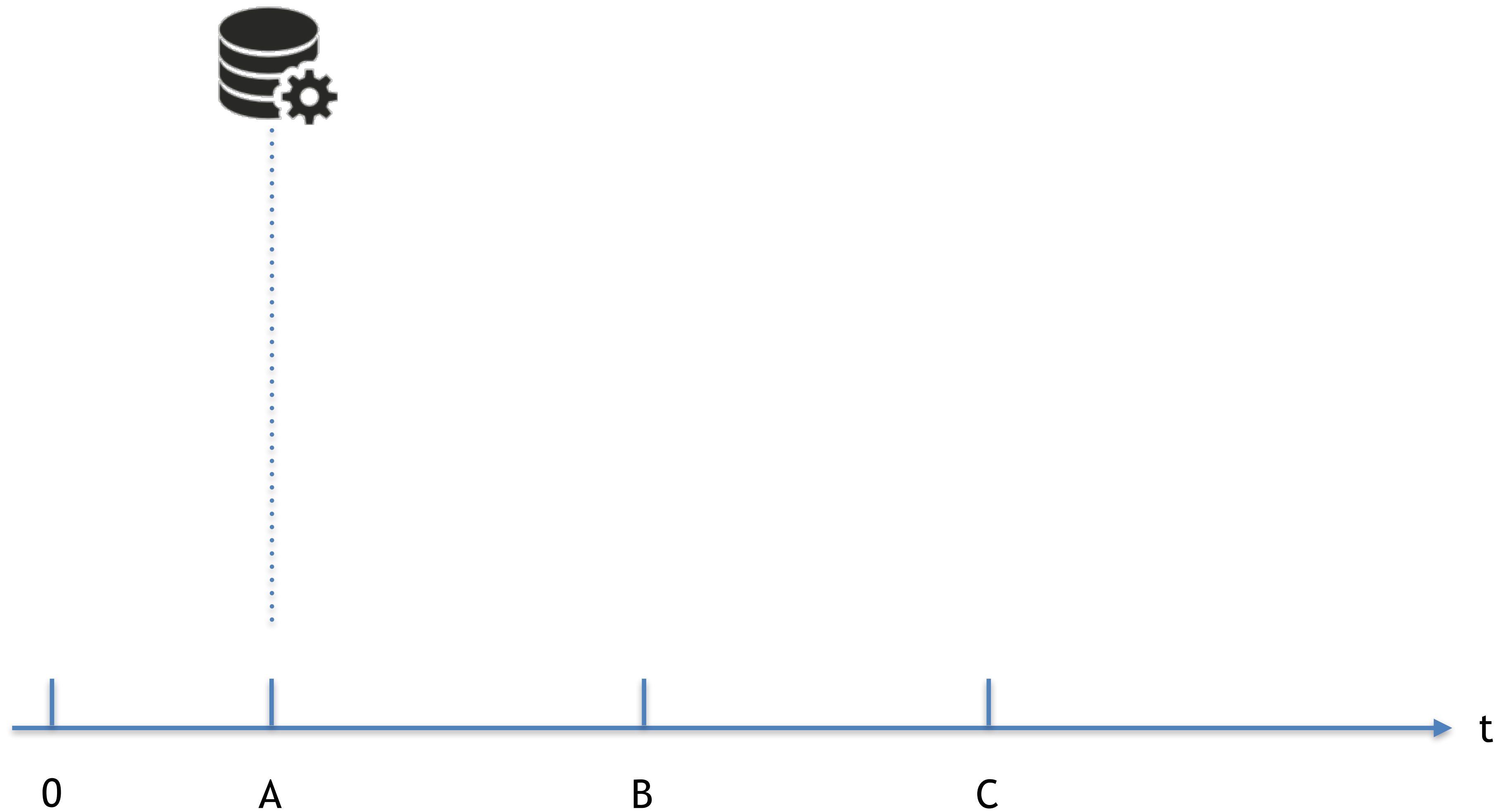


Более умное решение

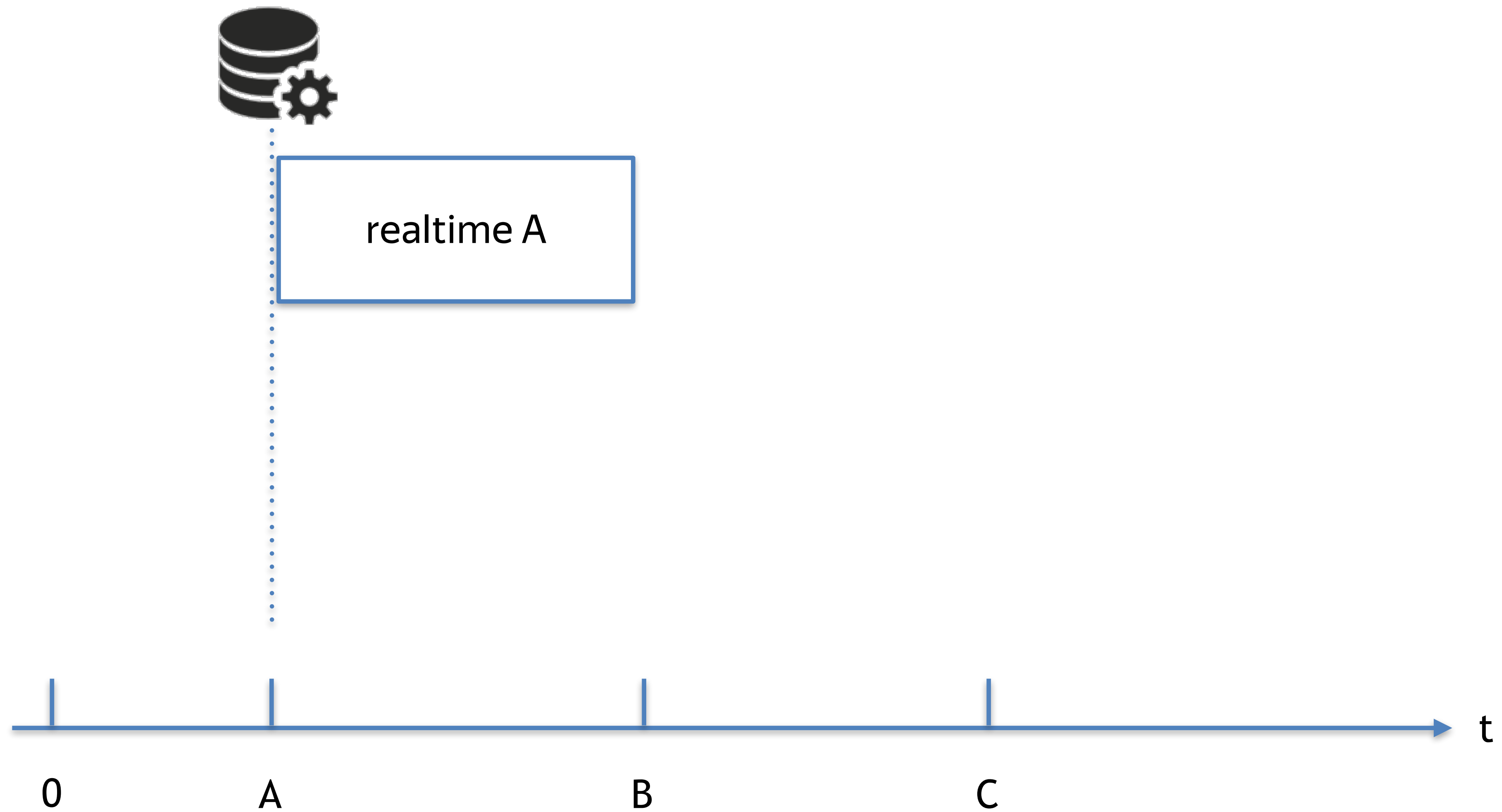
Более умное решение



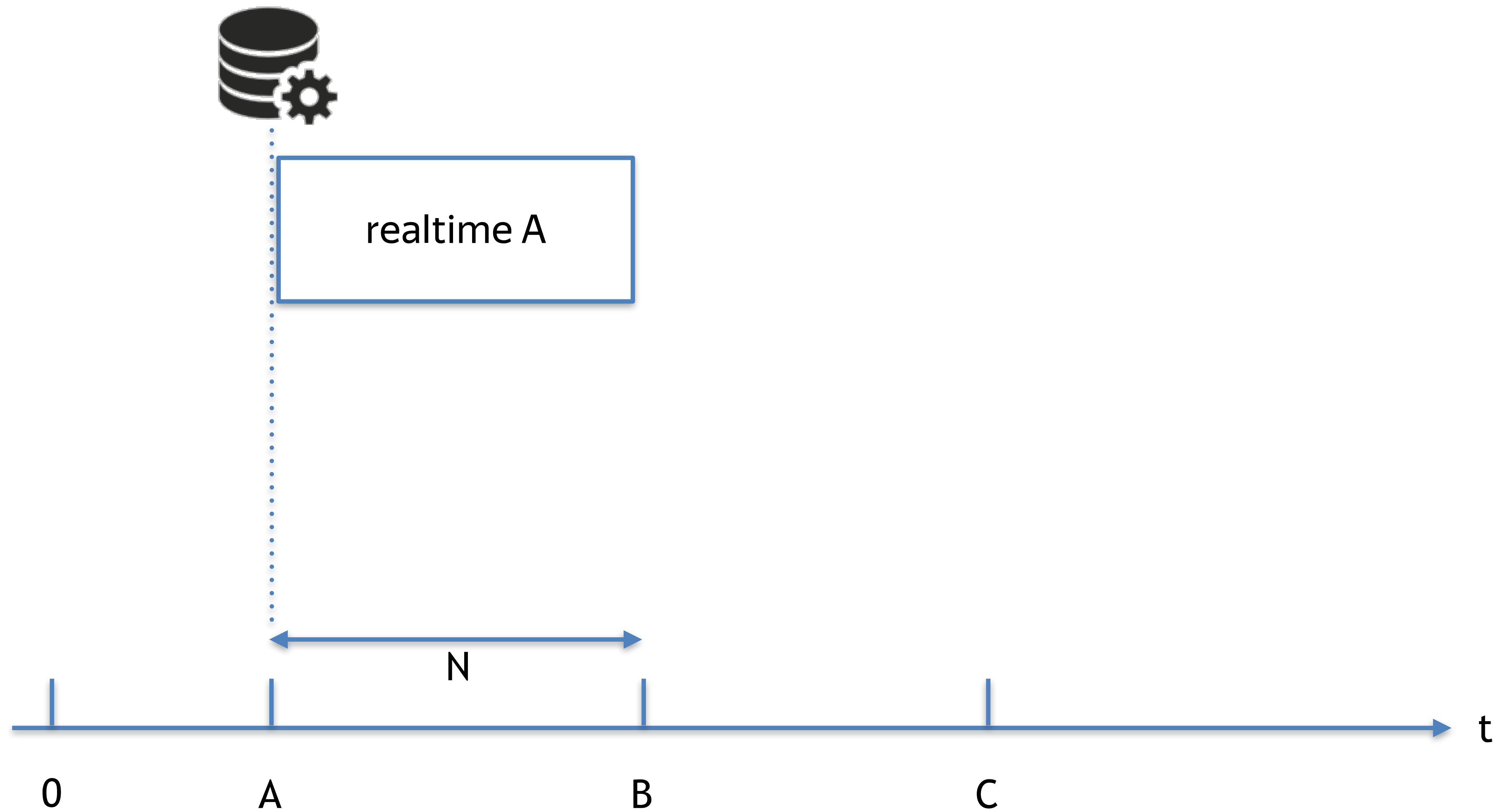
Более умное решение



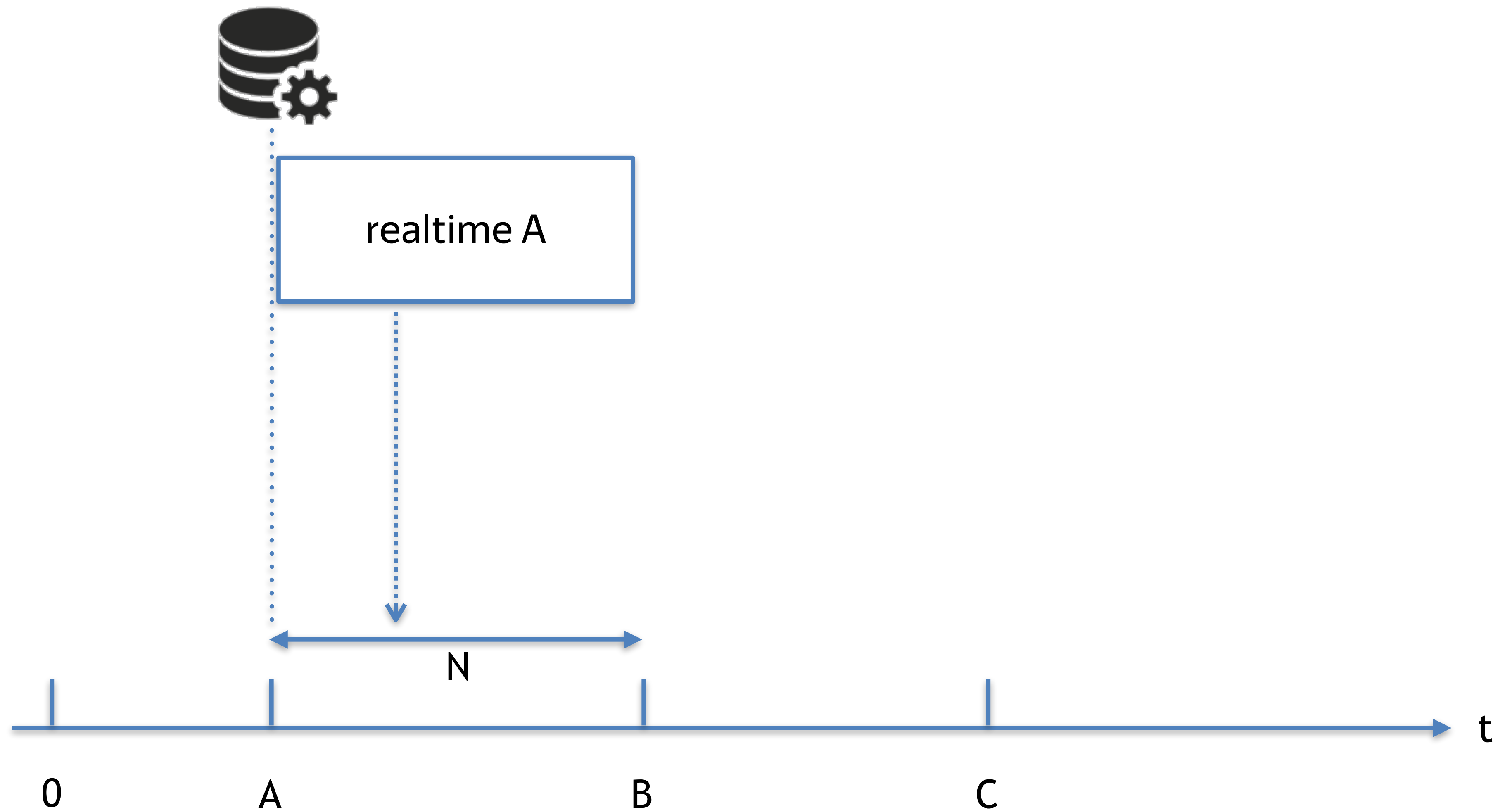
Более умное решение



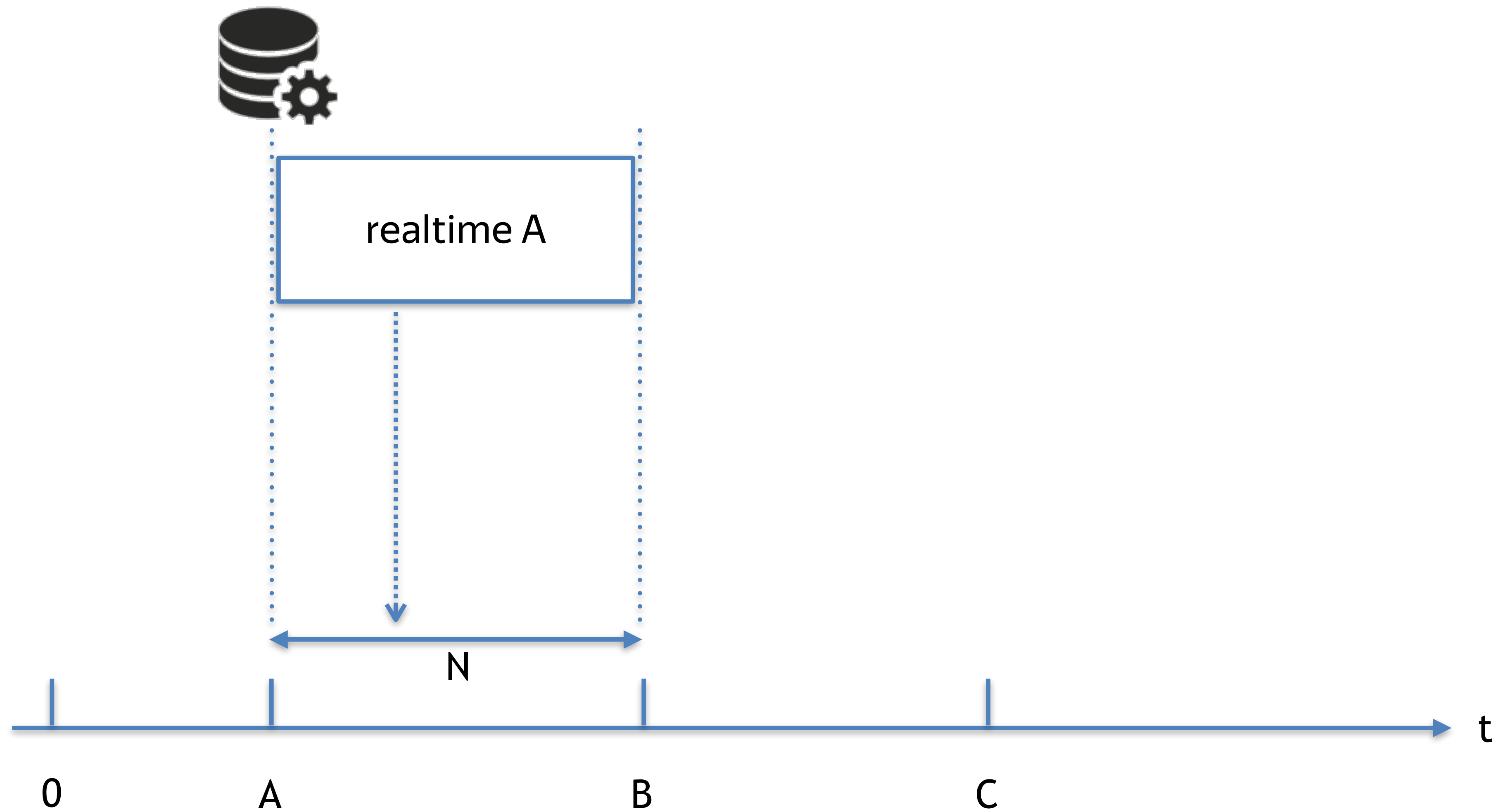
Более умное решение



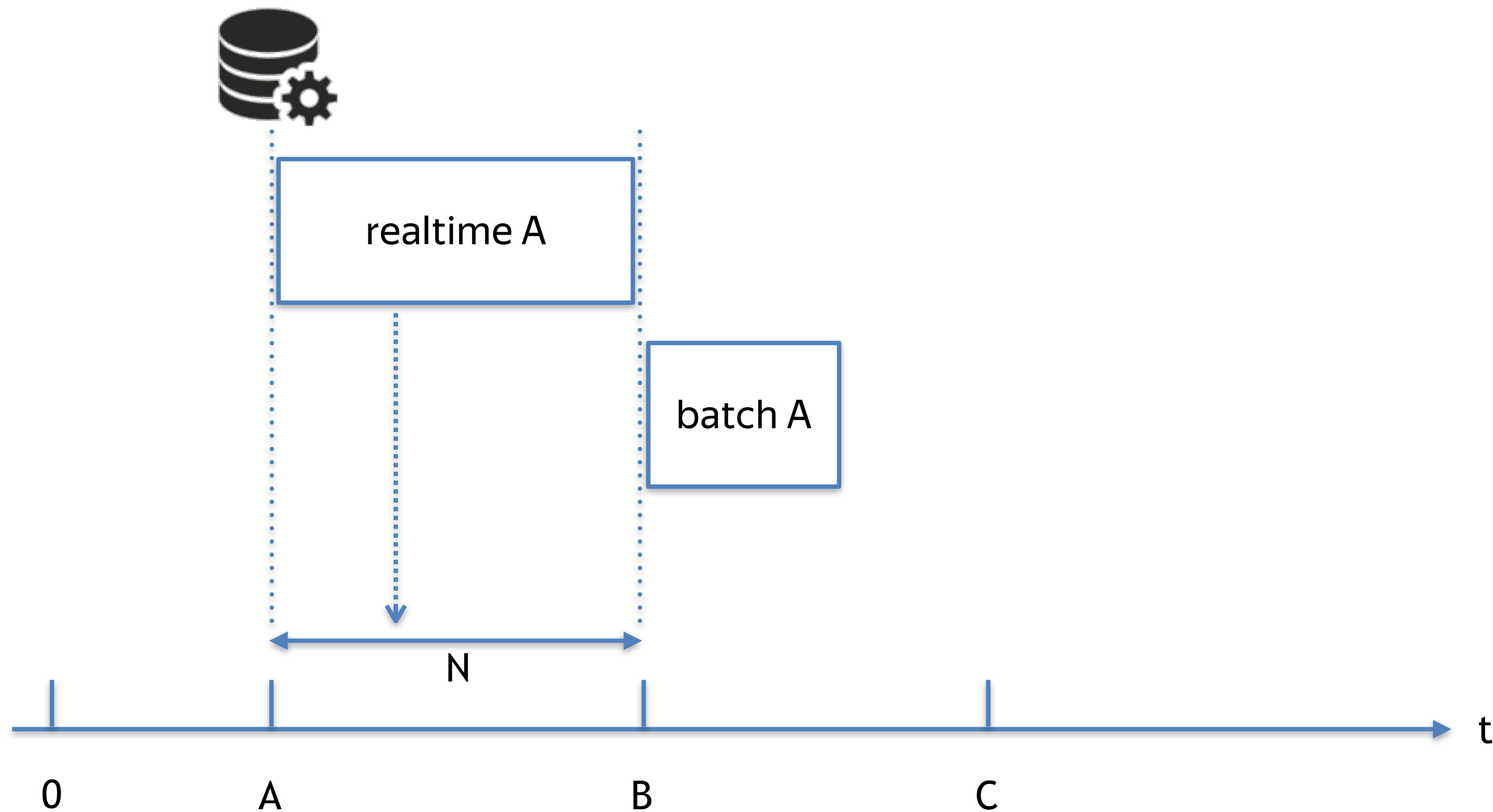
Более умное решение



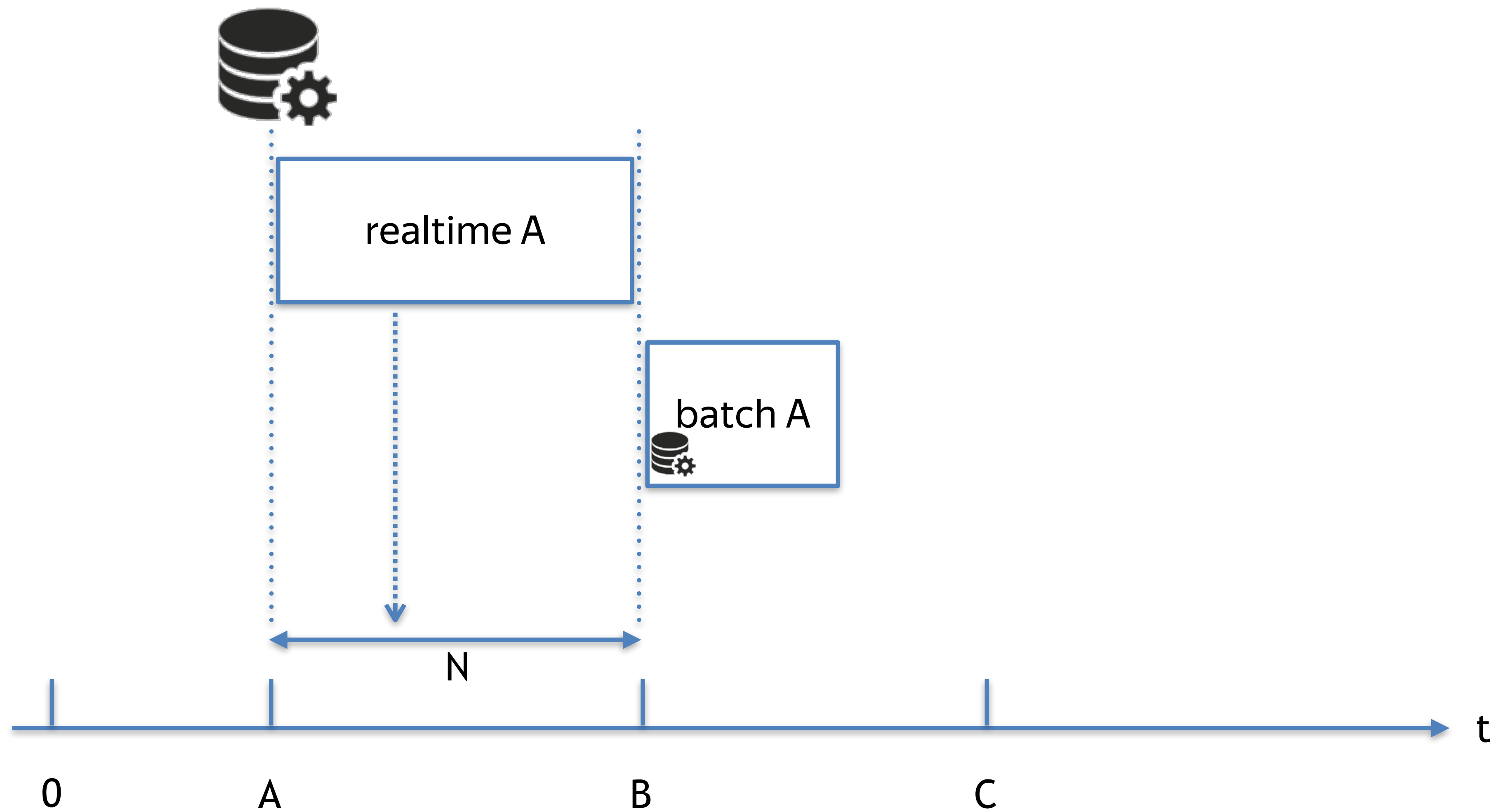
Более умное решение



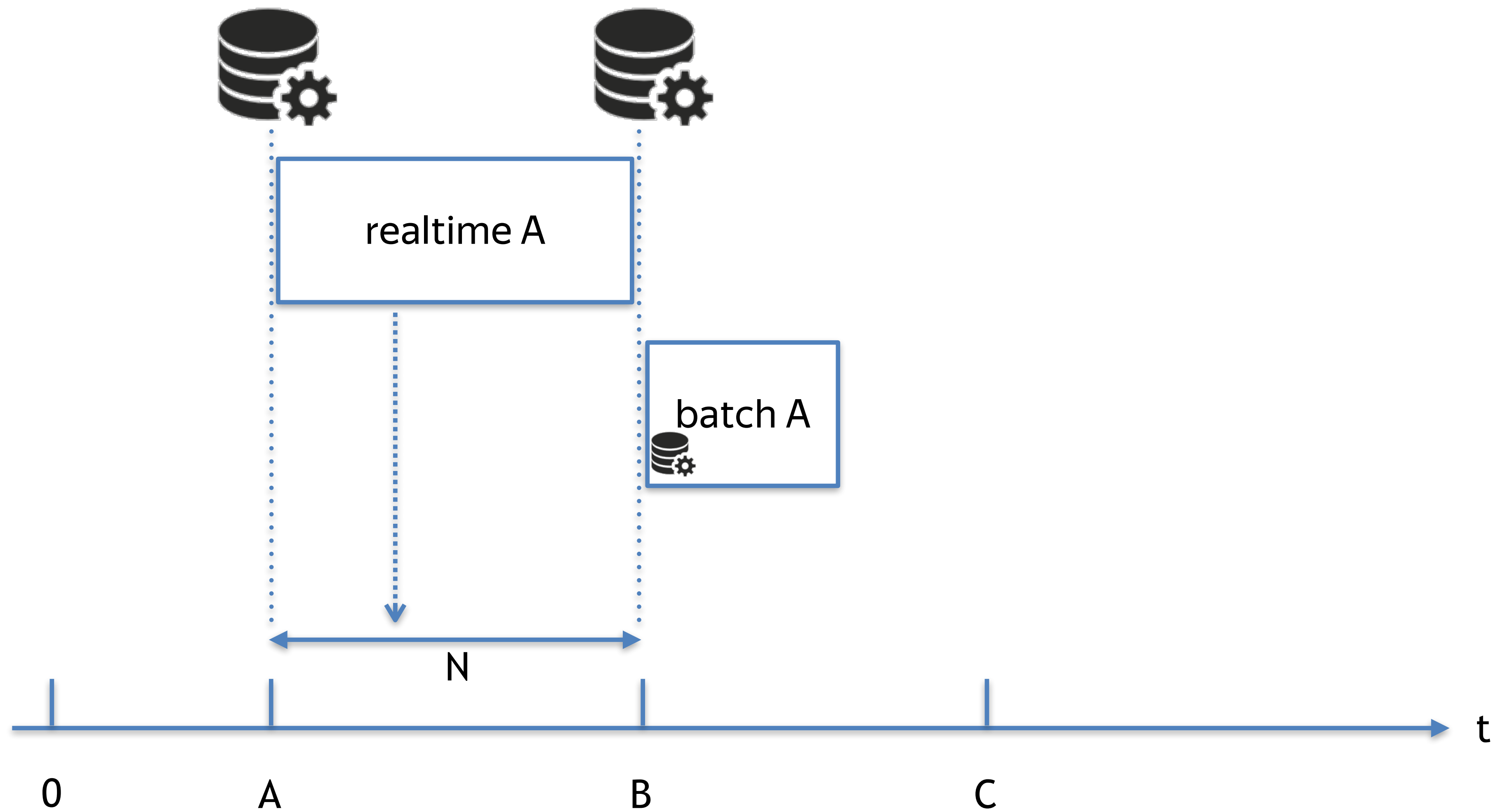
Более умное решение



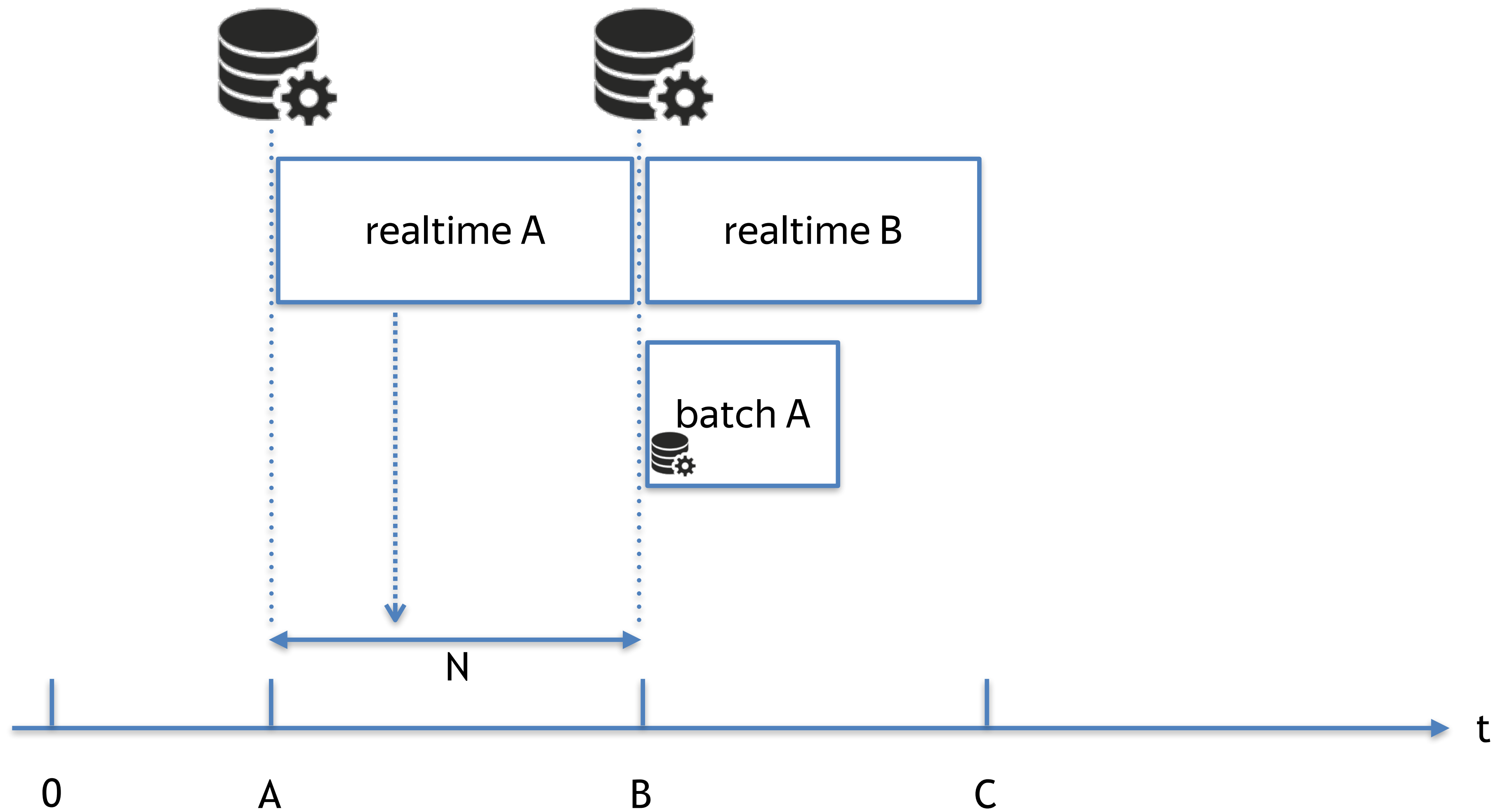
Более умное решение



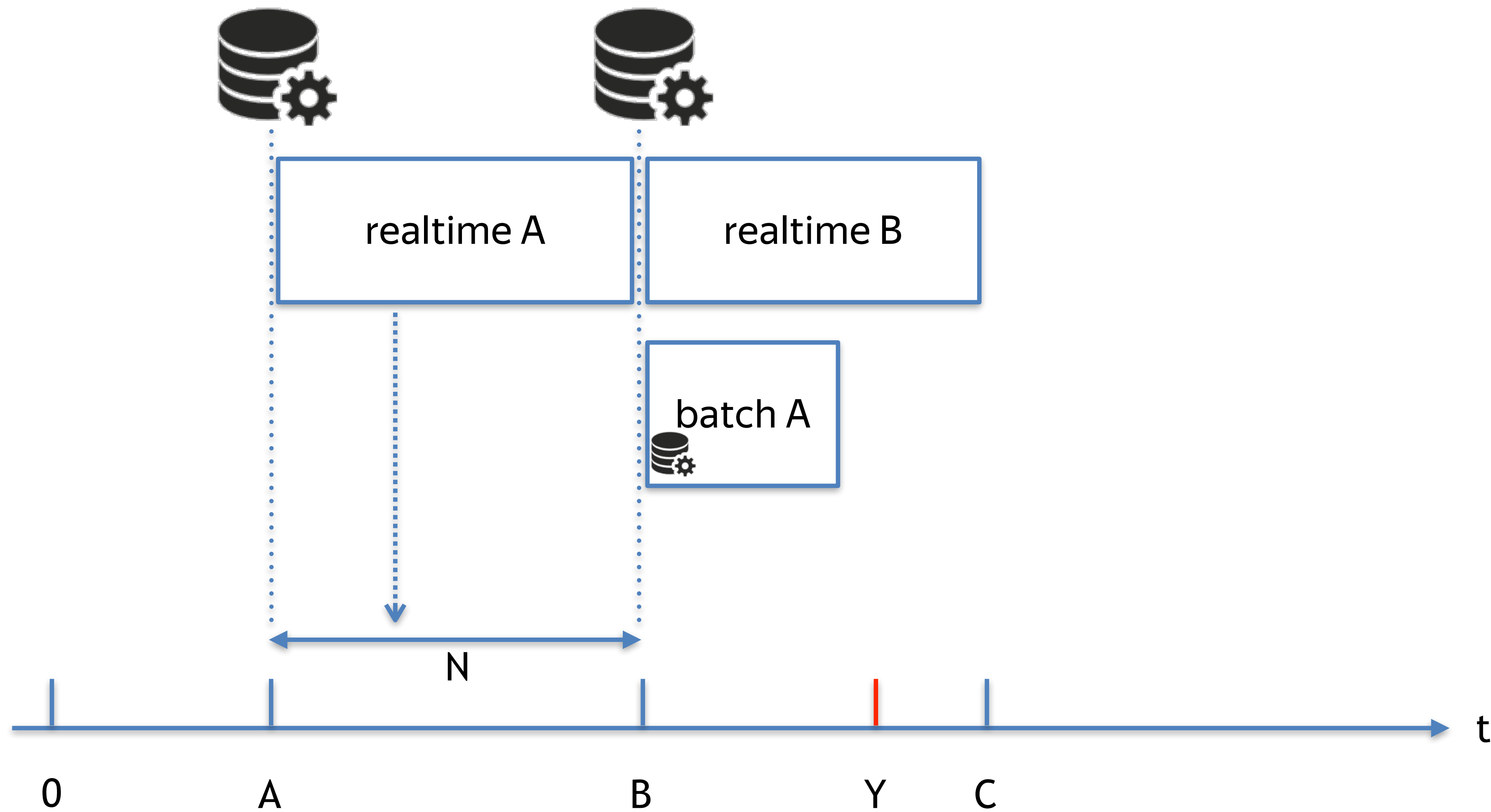
Более умное решение



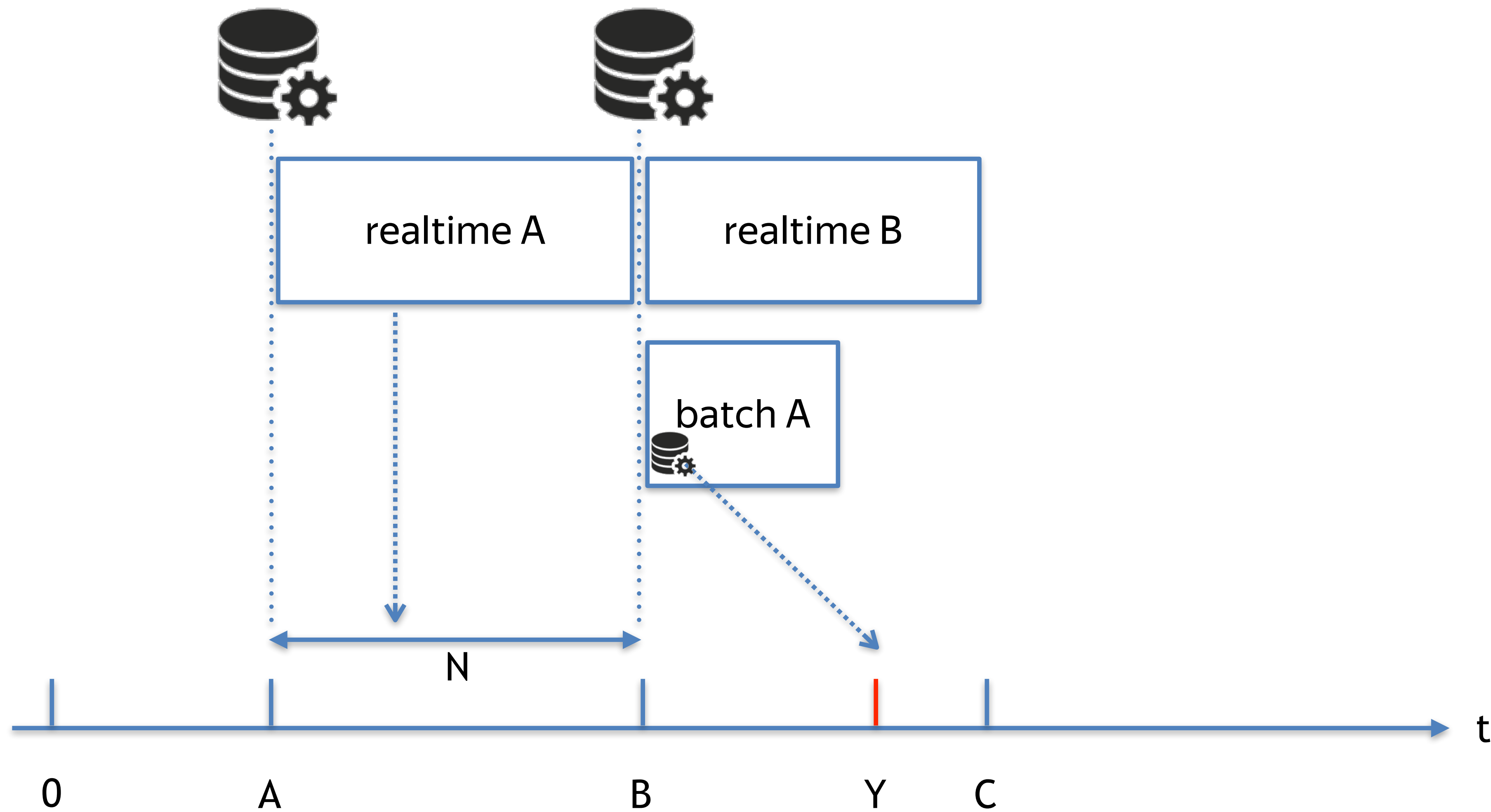
Более умное решение



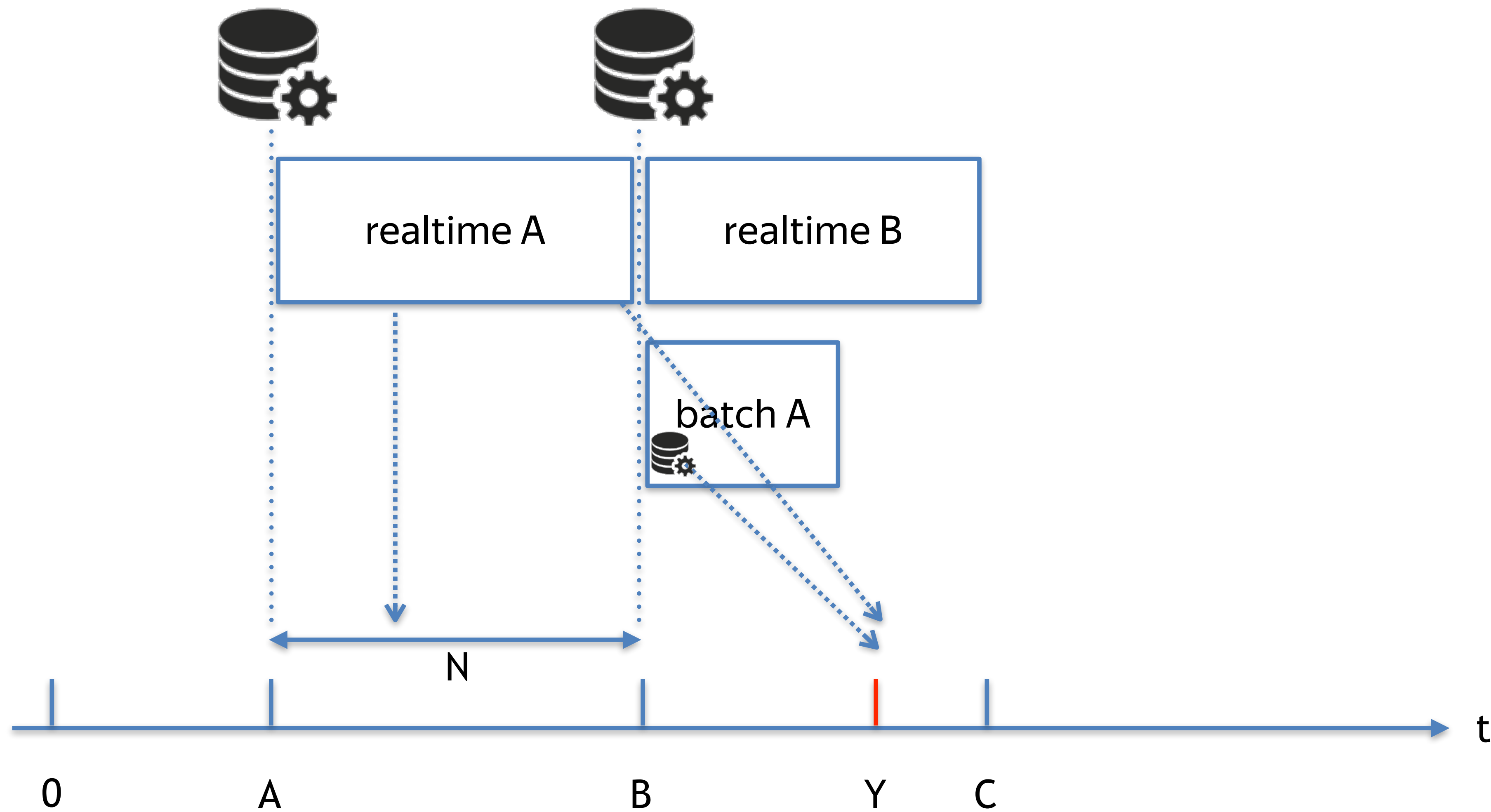
Более умное решение



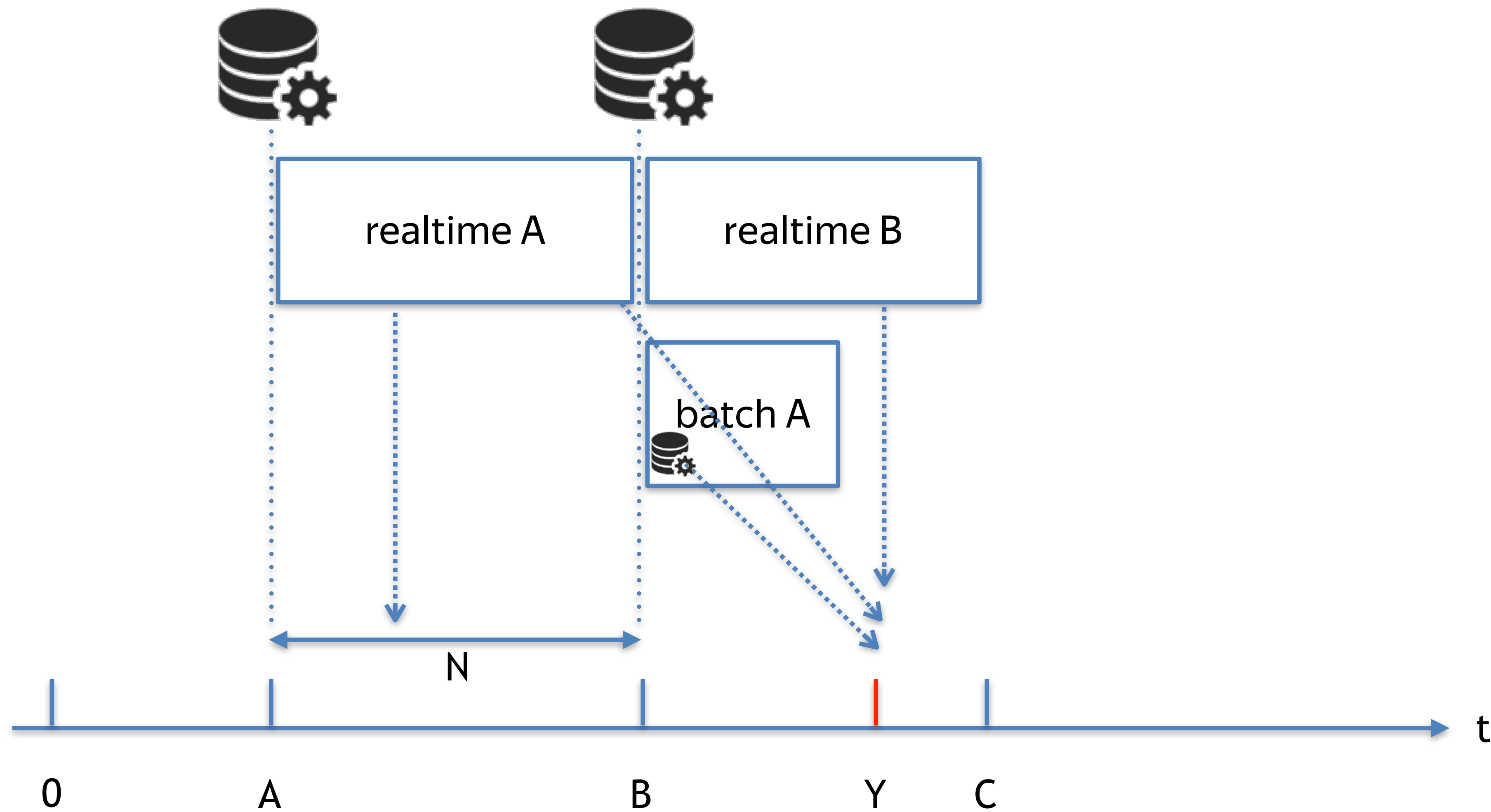
Более умное решение



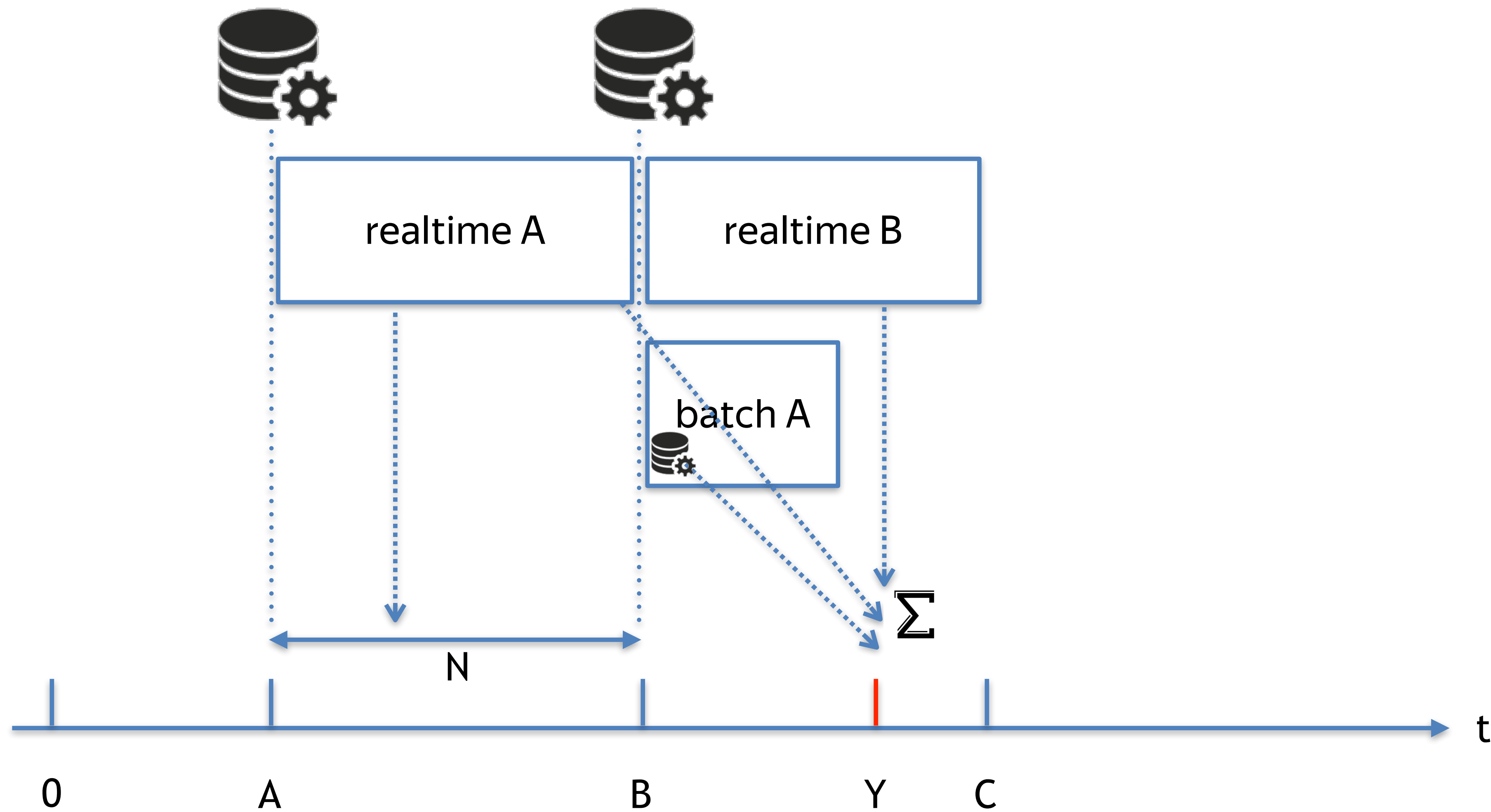
Более умное решение



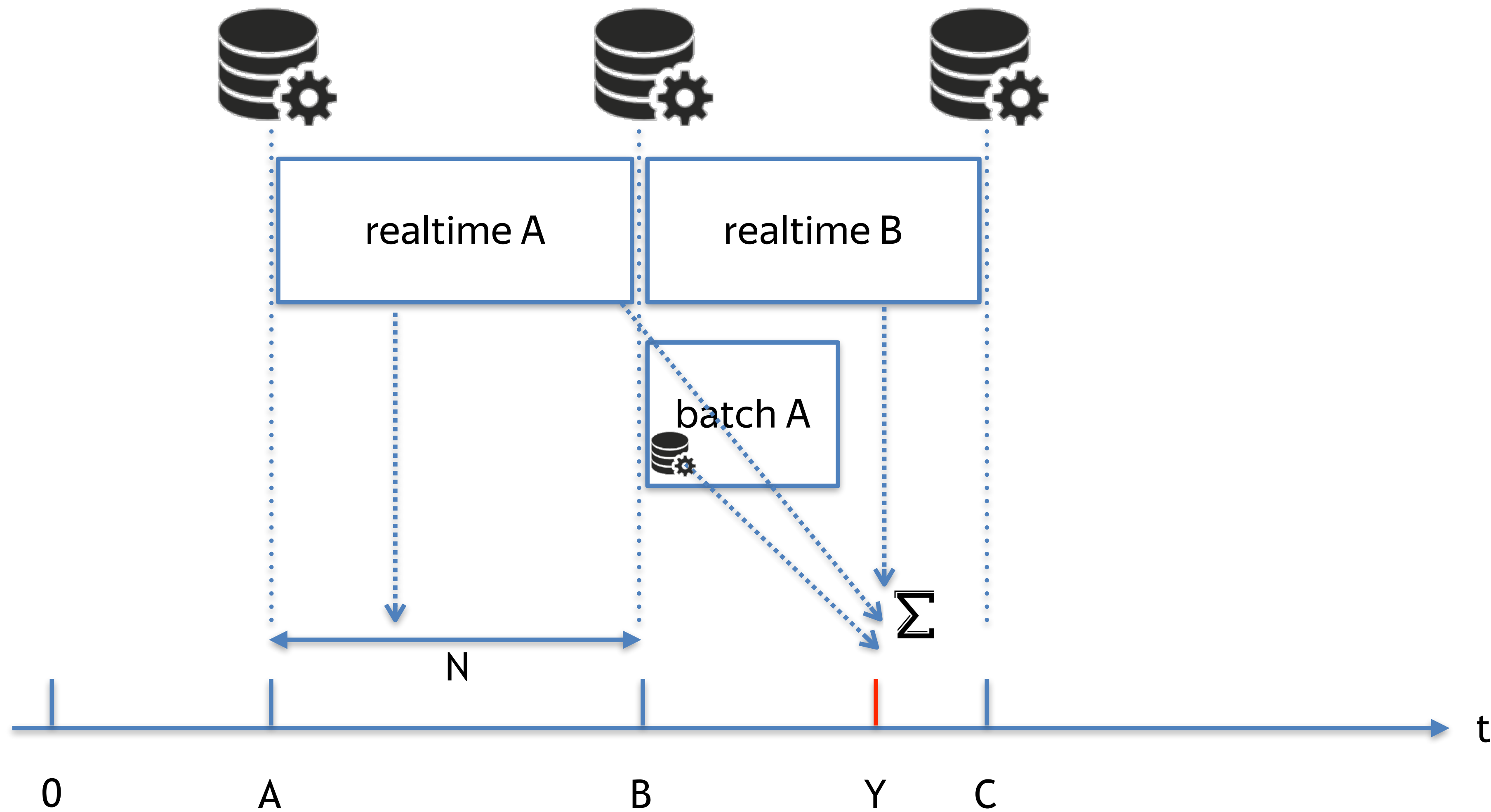
Более умное решение



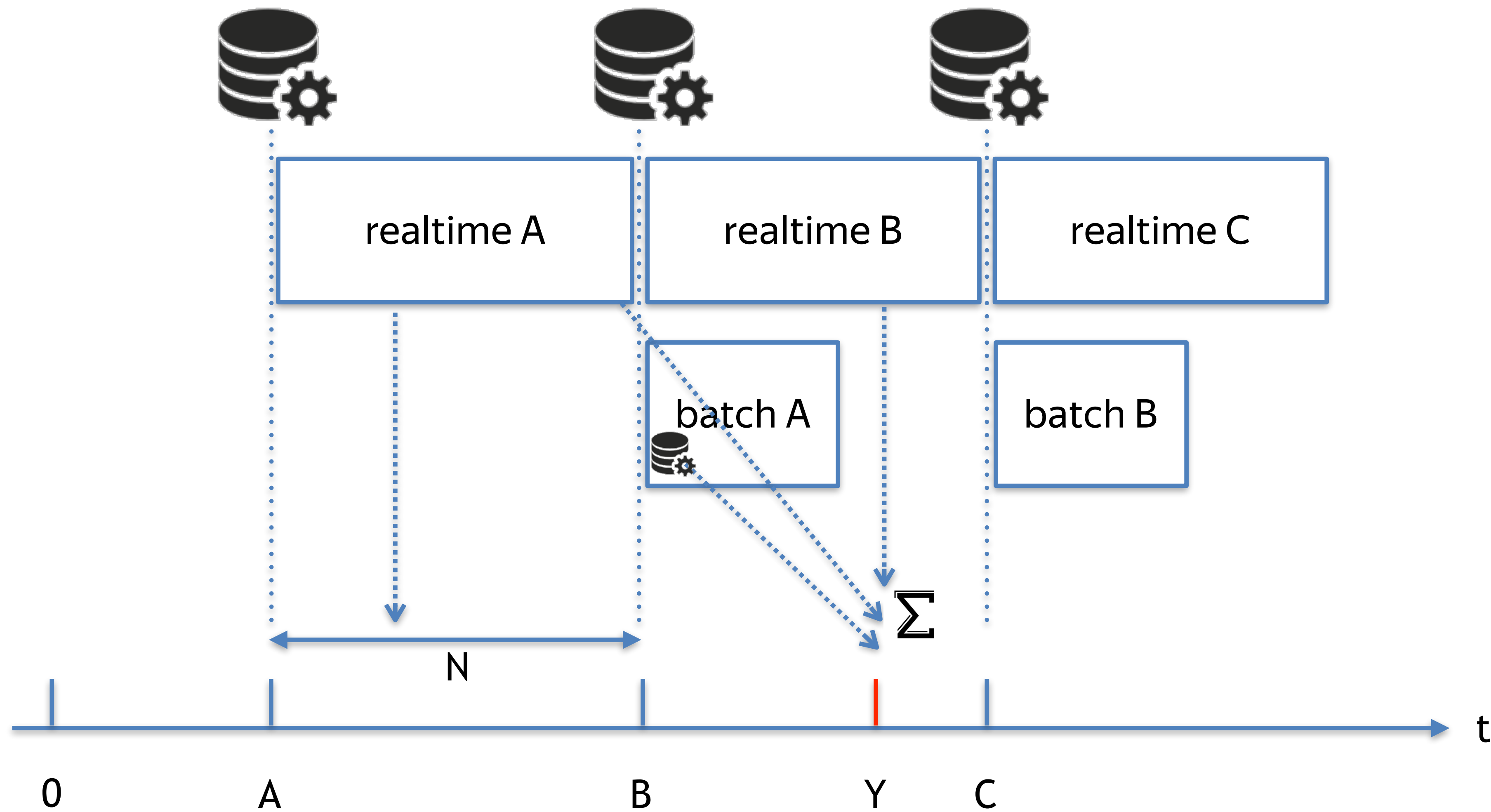
Более умное решение



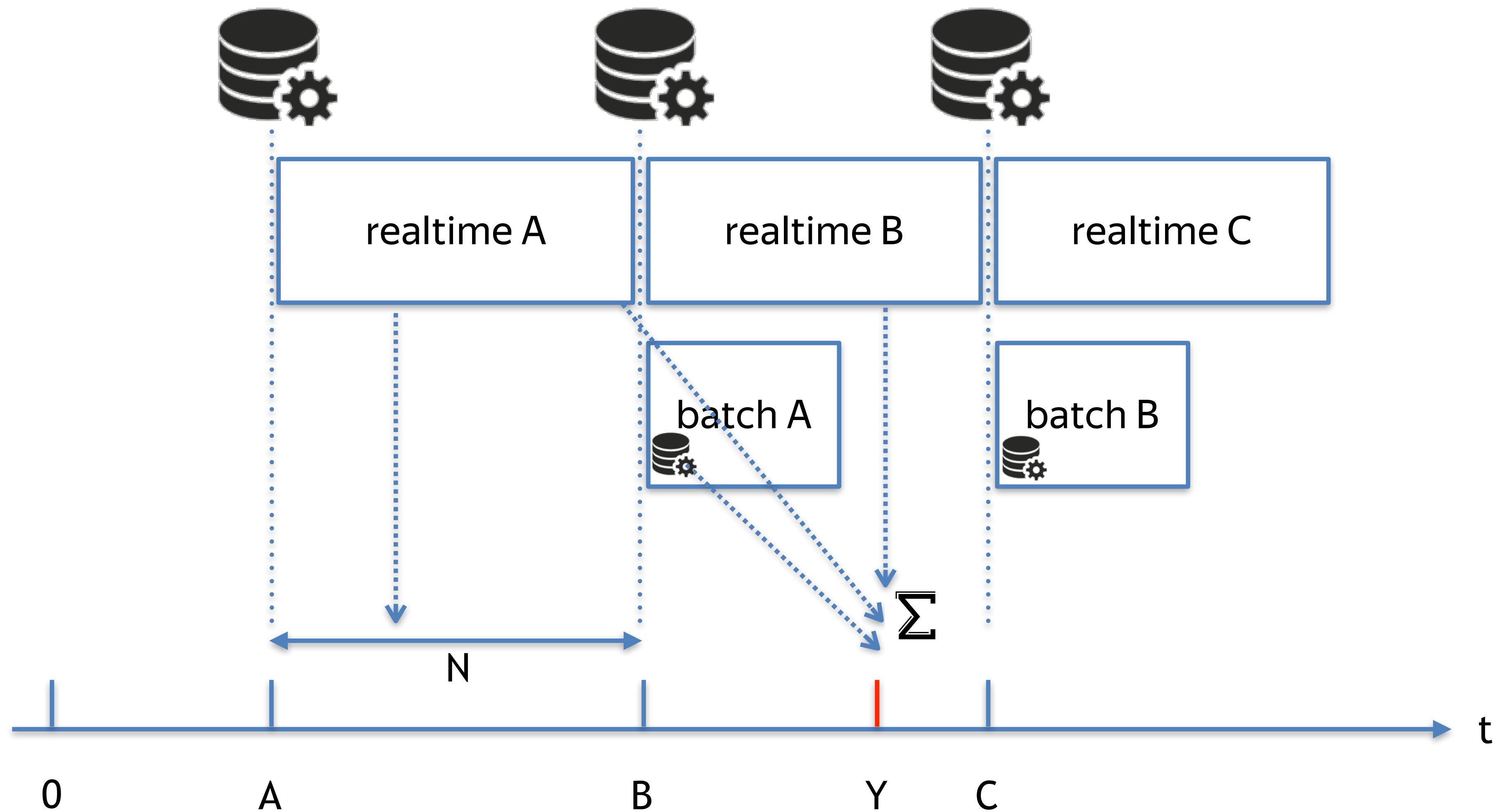
Более умное решение



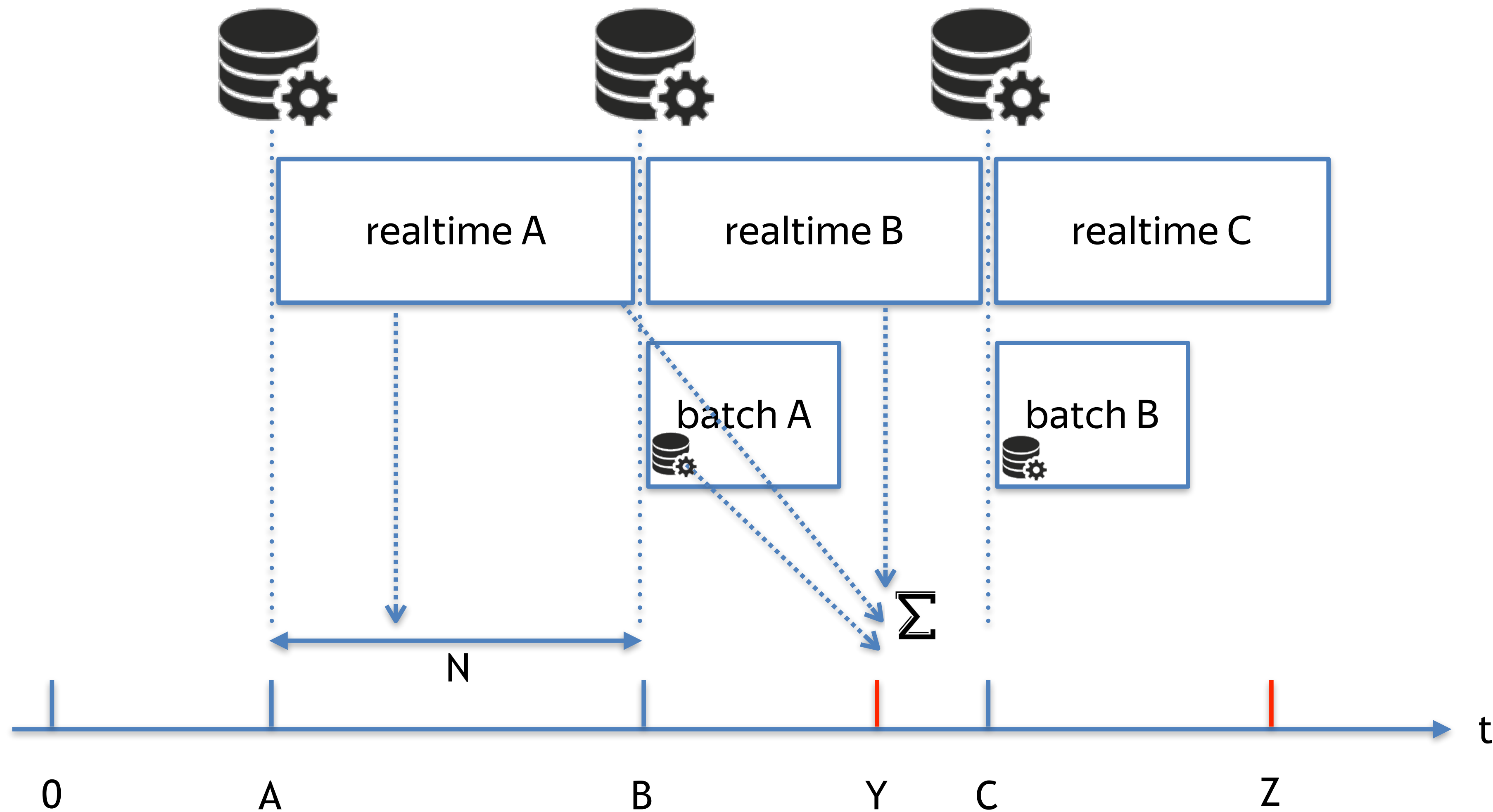
Более умное решение



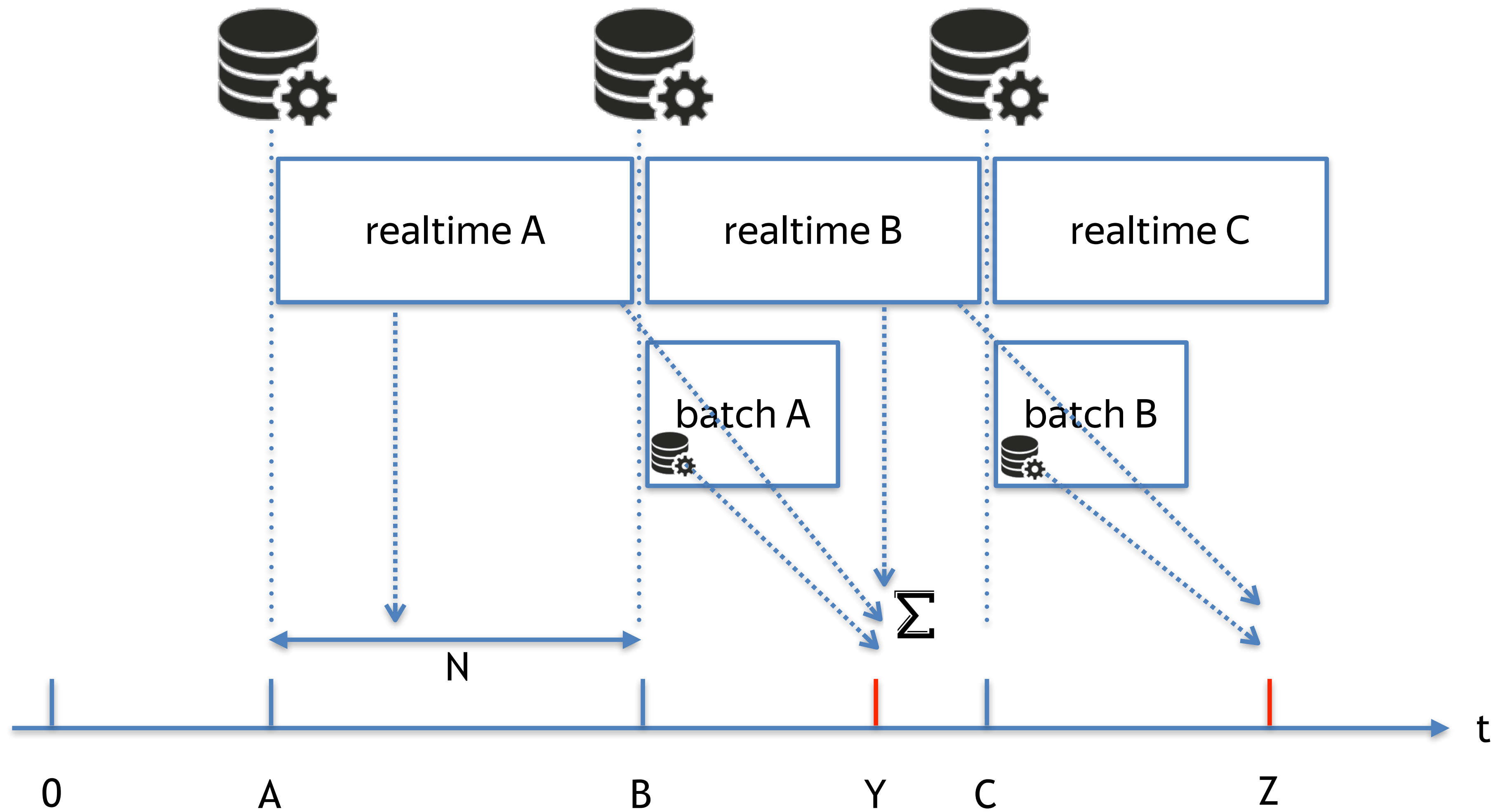
Более умное решение



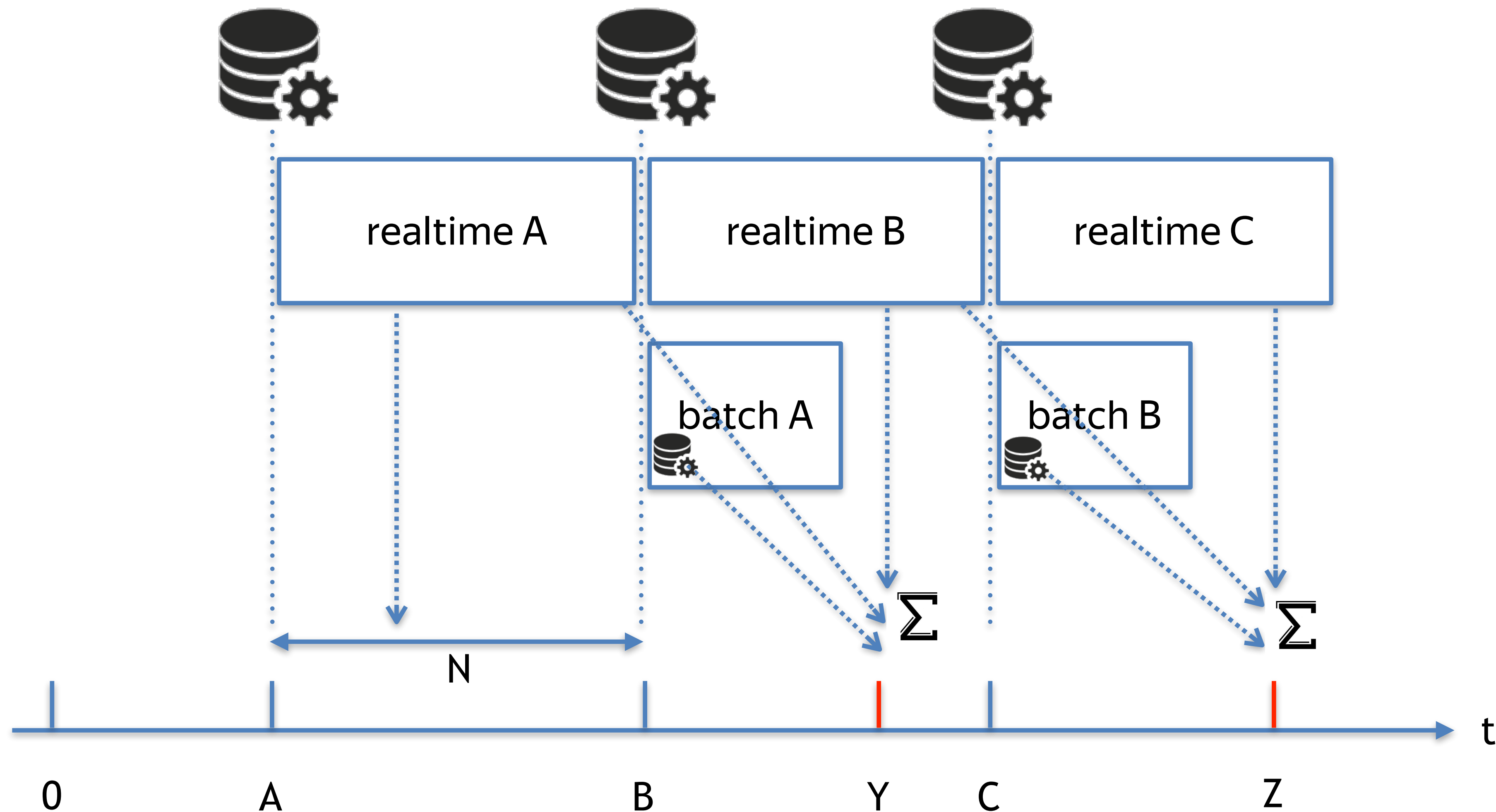
Более умное решение



Более умное решение

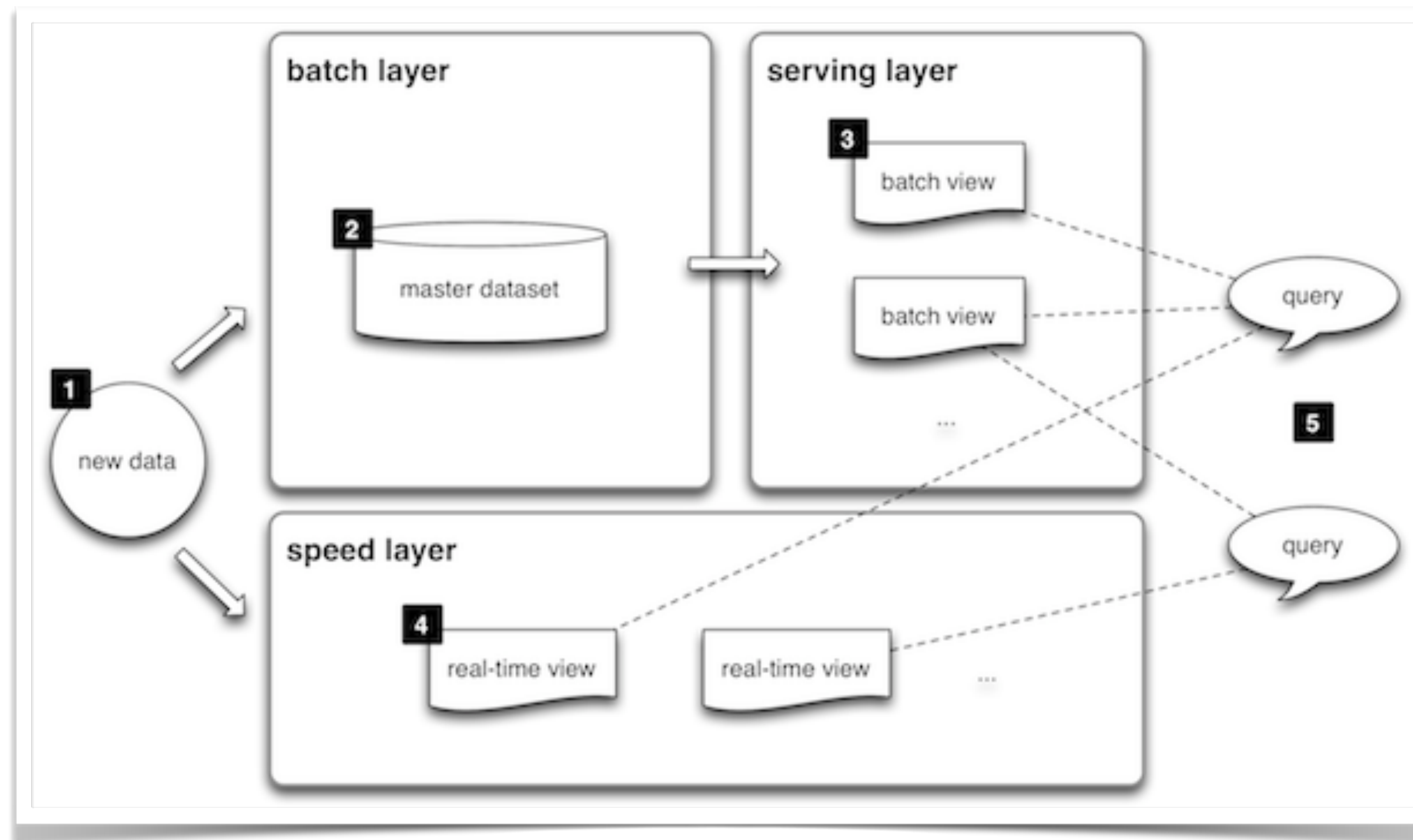


Более умное решение



Лямбда-архитектура (Nathan Marz)

<http://lambda-architecture.net>



*Realtime

Быстрые данные (Fast data): время от момента получения данных до момента выделения из них ценной информации

Big Data – Определение #1

Big Data – Определение #1

Много данных как результат:

- > избыточности
- > иммутабельности
- > репликации

Какие технологии подходят для
реализации master dataset?



Файловые хранилища

Файловые хранилища

- > HDFS
- > WebDAV
- > Amazon S3

NoSql базы данных

NoSql базы данных

документо-ориентированные (Mongo, MarkLogic)
и key-value (Cassandra, Hazelcast)

NoSql базы данных

документо-ориентированные (Mongo, MarkLogic)
и key-value (Cassandra, Hazelcast)

распределенные

PostgreSQL не подходит для
реализации master dataset!



PostgreSQL **тоже** подходит для
реализации master dataset!



Big Data – Определение #2

Big Data – Определение #2

3 технологических признака:

Big Data – Определение #2

3 технологических признака:

1) Гибкое хранилище и возможность делать разнообразные запросы (Mongo, Cassandra, HDFS, Postgres)

Каскадное хранилище

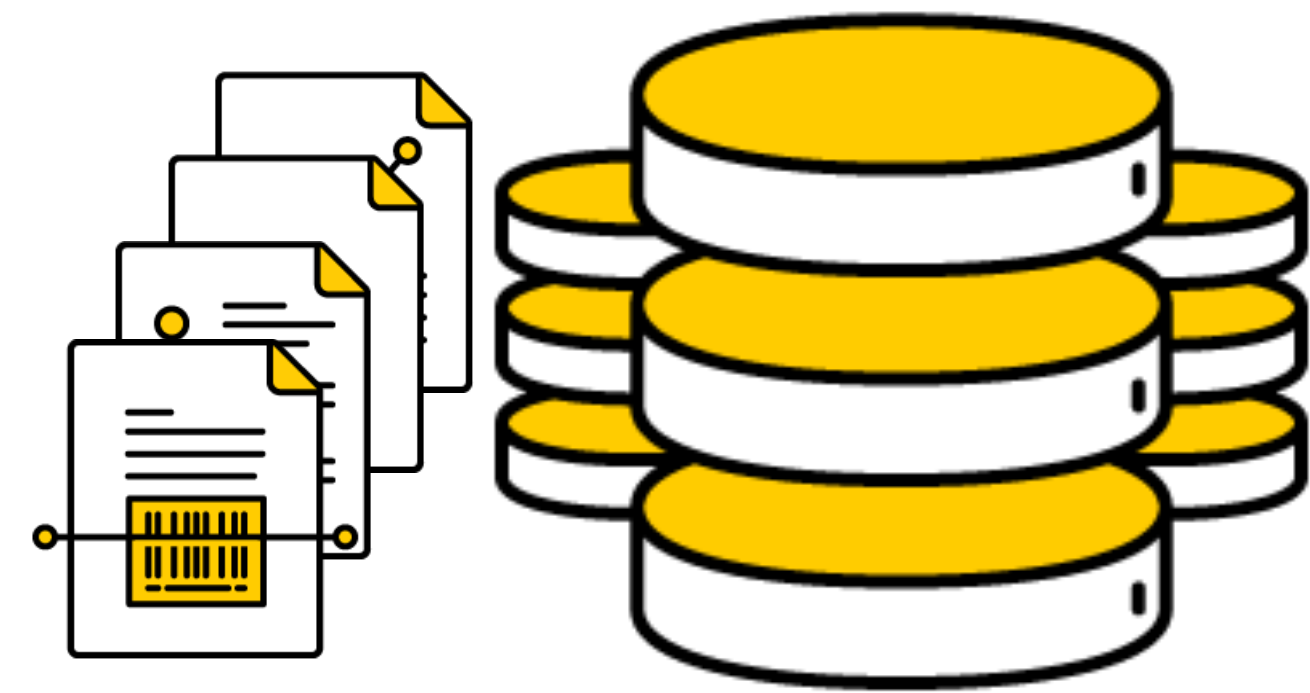
Каскадное хранилище

«Резиновая помойка»



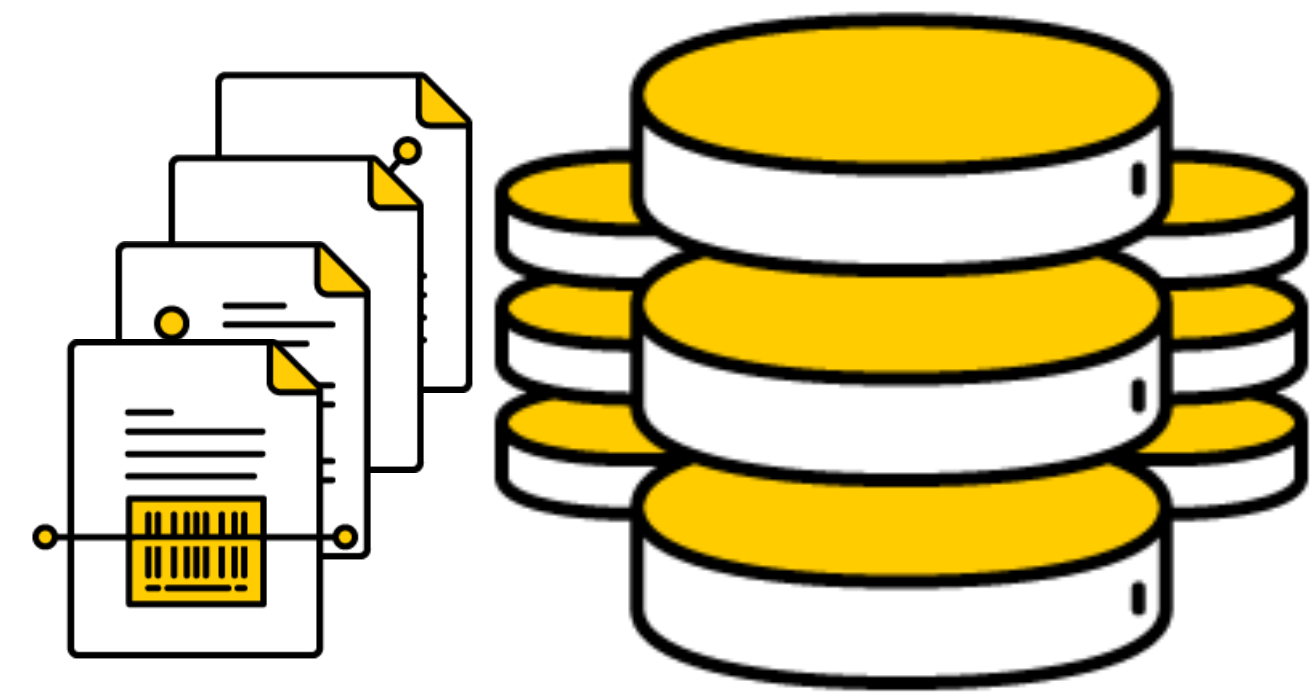
Каскадное хранилище

«Резиновая помойка»



Каскадное хранилище

«Резиновая помойка»



HDFS, S3, WebDAV

Каскадное хранилище

«Резиновая помойка»

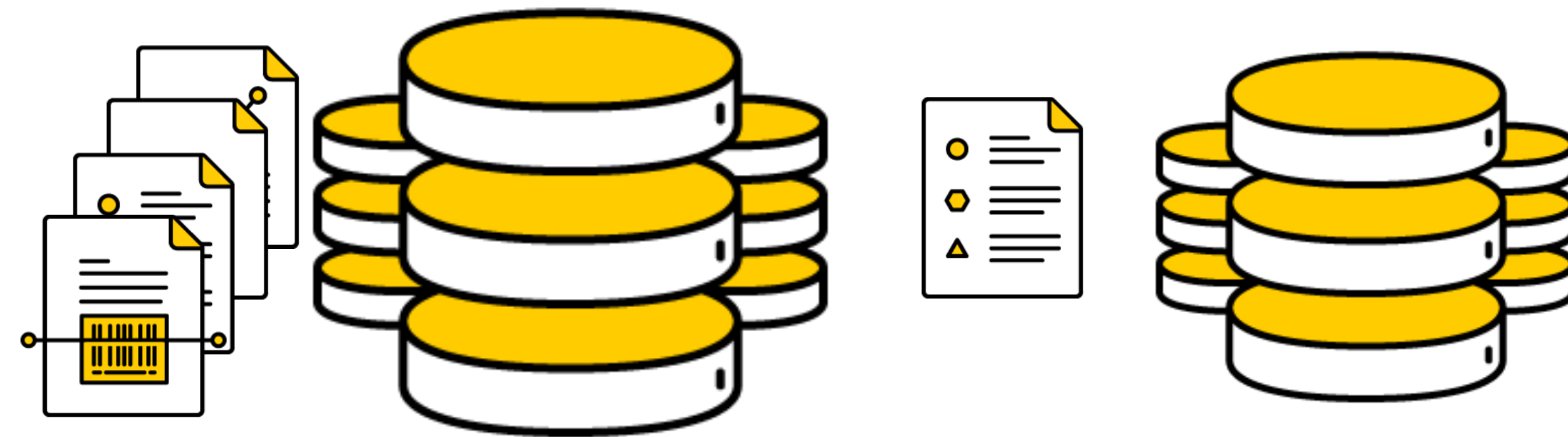


HDFS, S3, WebDAV

Каскадное хранилище

«Резиновая помойка»

Хранилище фактов

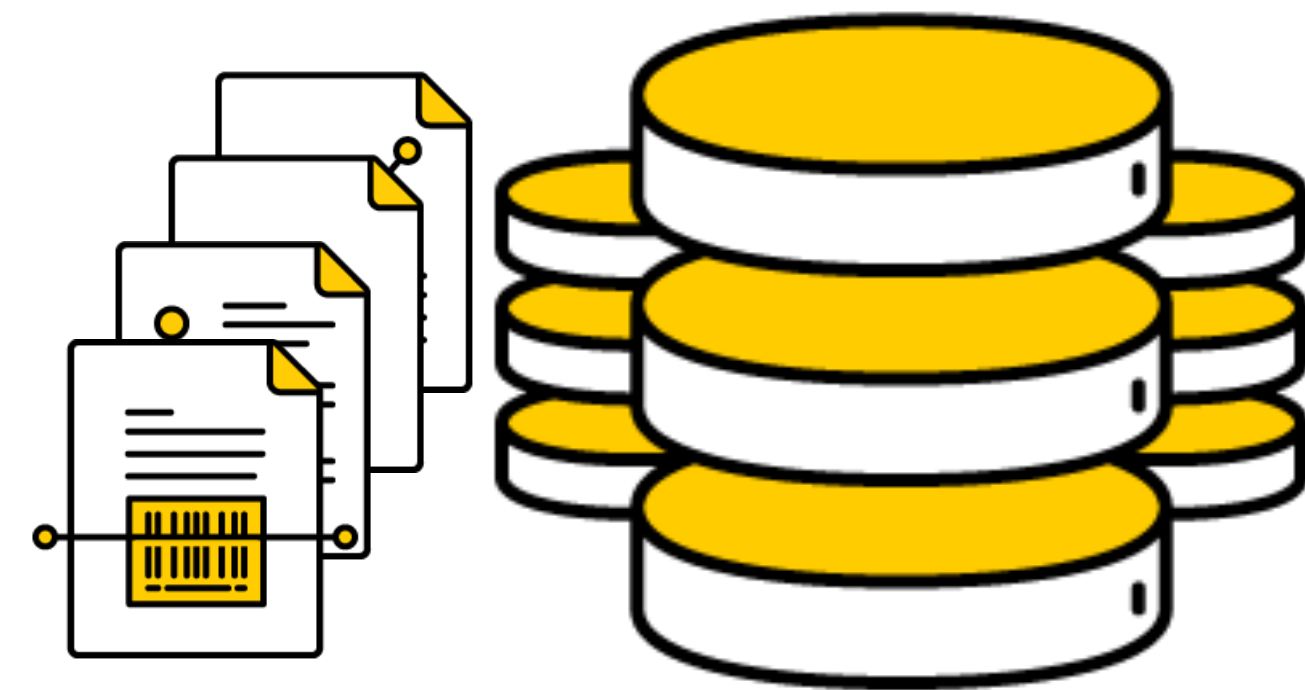


HDFS, S3, WebDAV

Каскадное хранилище

«Резиновая помойка»

Хранилище фактов



HDFS, S3, WebDAV

Mongo, MarkLogic,
Postgres, Cassandra

Каскадное хранилище

«Резиновая помойка»

Хранилище фактов



HDFS, S3, WebDAV

Mongo, MarkLogic,
Postgres, Cassandra

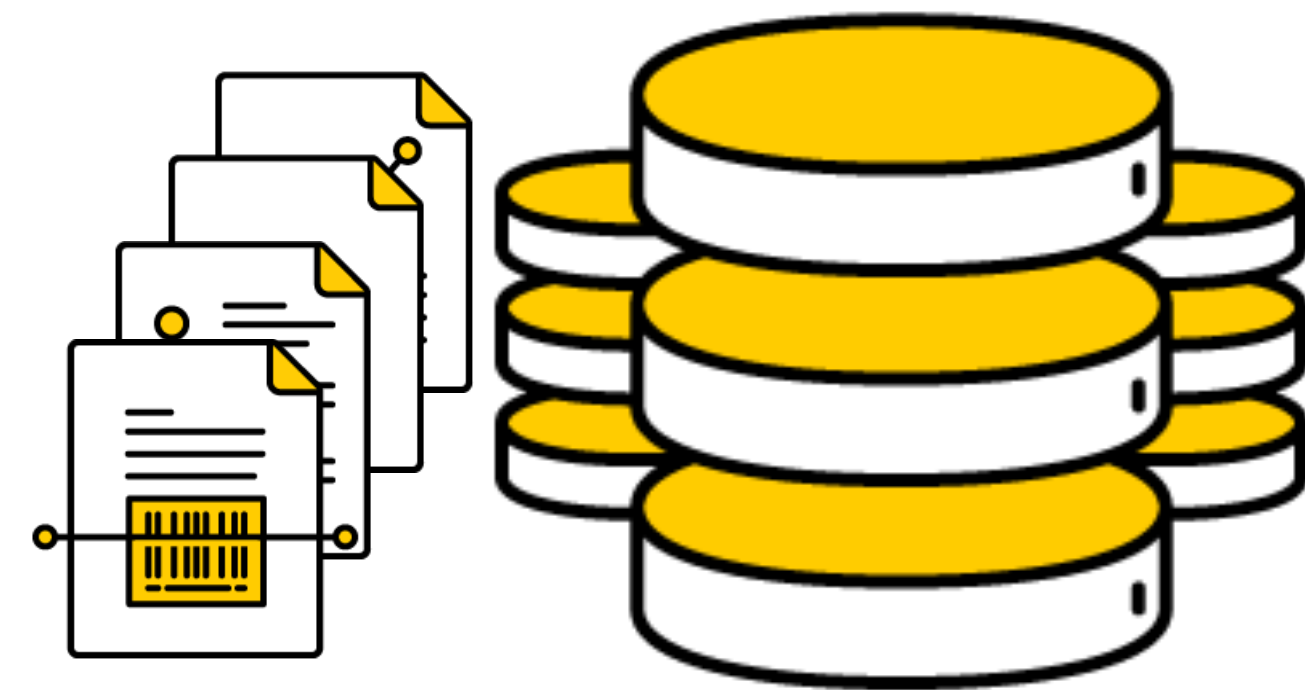
Каскадное хранилище

«Резиновая помойка»

Хранилище фактов



SQL databases: Oracle, Postgres, MySql



HDFS, S3, WebDAV



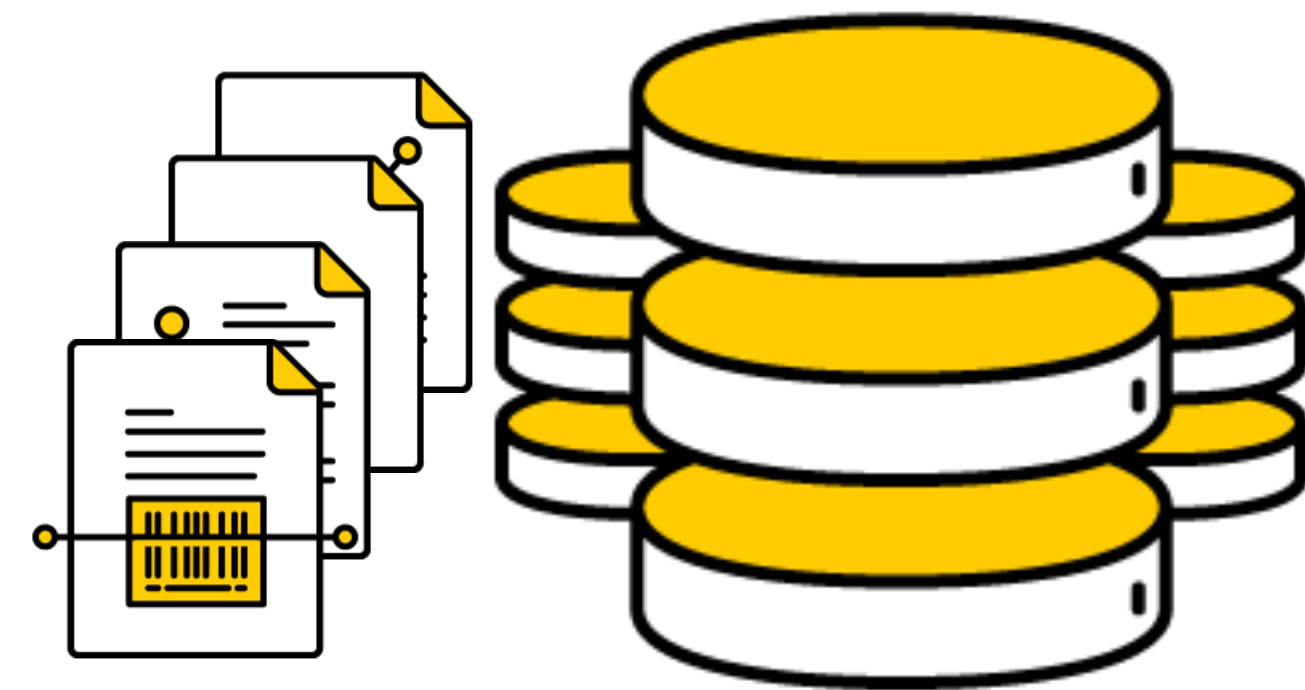
Mongo, MarkLogic, Postgres, Cassandra

Каскадное хранилище

«Резиновая помойка»

Хранилище фактов

SQL databases: Oracle, Postgres, MySql

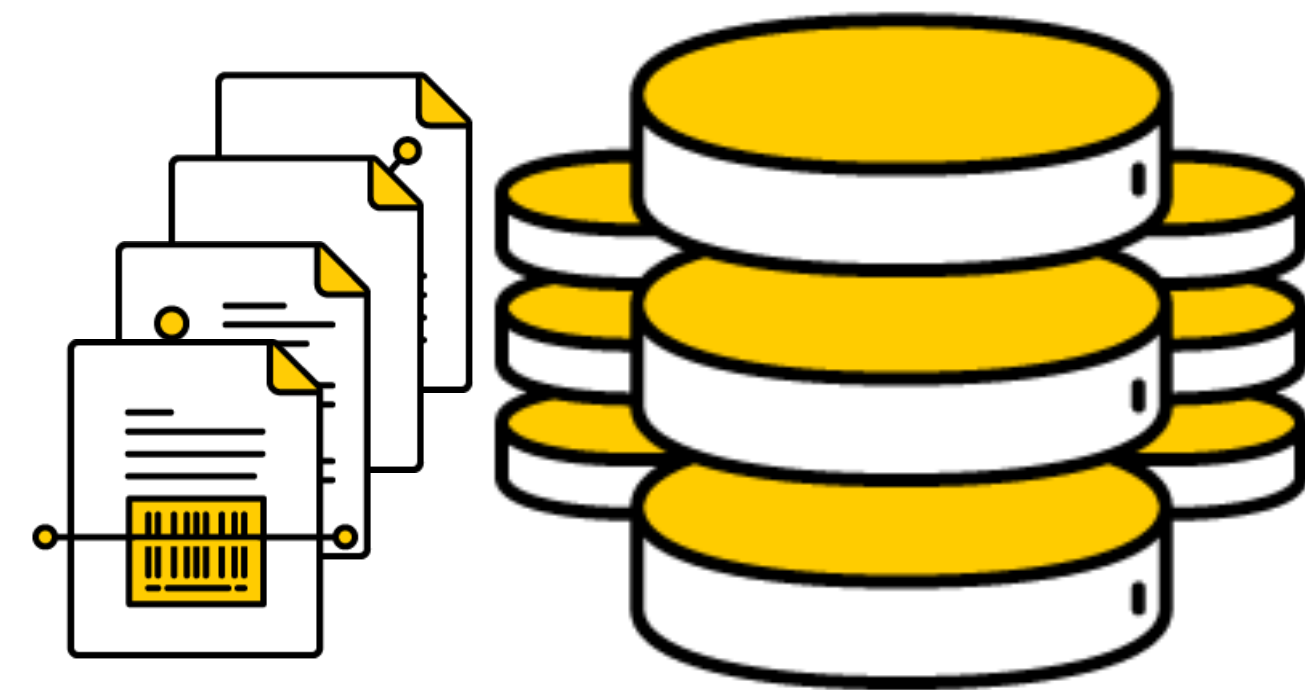


HDFS, S3, WebDAV

Mongo, MarkLogic, Postgres, Cassandra

Каскадное хранилище

«Резиновая помойка»

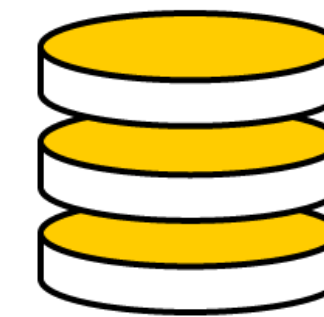


HDFS, S3, WebDAV

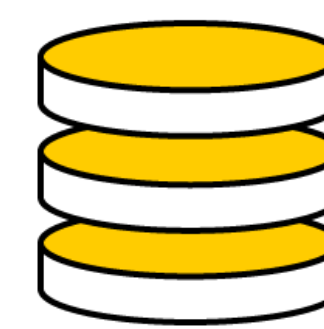
Хранилище фактов



Mongo, MarkLogic,
Postgres, Cassandra



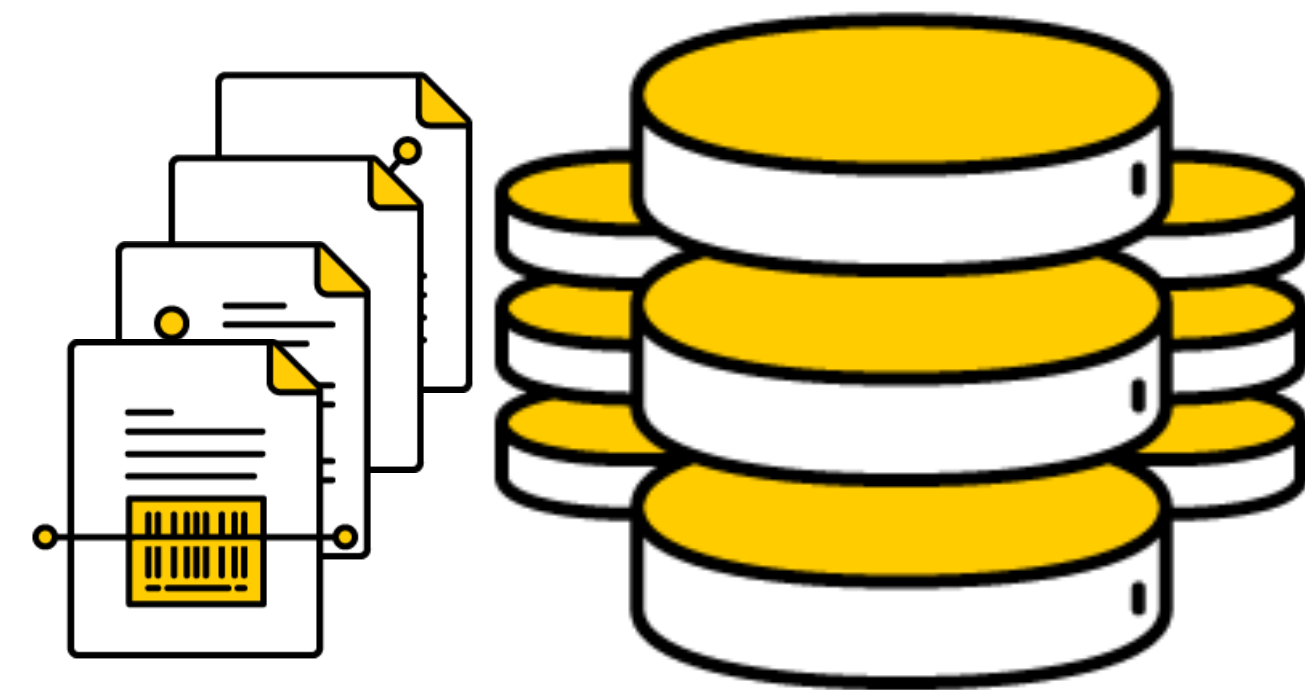
SQL databases: Oracle,
Postgres, MySql



NoSql databases: Parquet,
ClickHouse, YT

Каскадное хранилище

«Резиновая помойка»

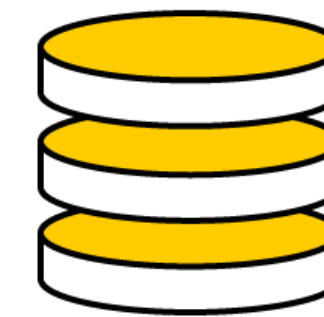


HDFS, S3, WebDAV

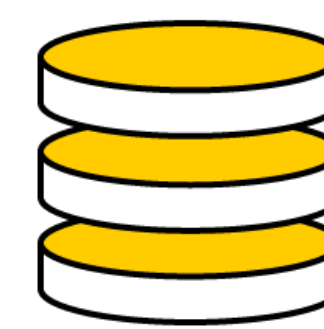
Хранилище фактов



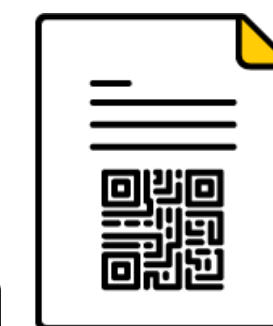
Mongo, MarkLogic,
Postgres, Cassandra



SQL databases: Oracle,
Postgres, MySql

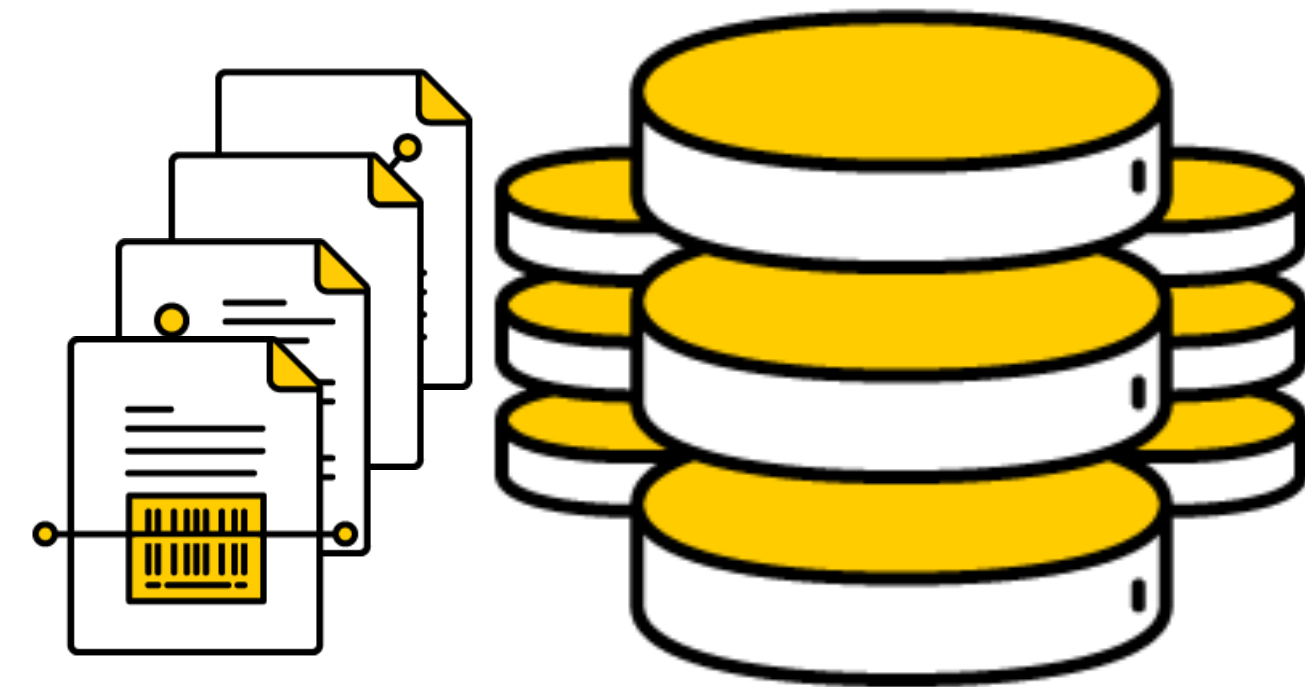


NoSql databases: Parquet,
ClickHouse, YT



Каскадное хранилище

«Резиновая помойка»

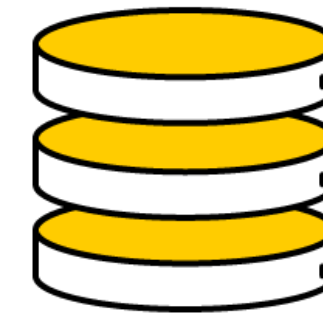


HDFS, S3, WebDAV

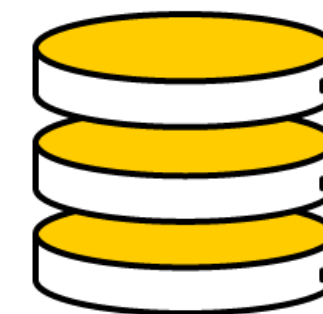
Хранилище фактов



Mongo, MarkLogic,
Postgres, Cassandra



SQL databases: Oracle,
Postgres, MySql



NoSql databases: Parquet,
ClickHouse, YT



Grids/Caches: Apache Ignite,
GridGain, Coherence,
Hazelcast

Про магазин и школьницу

<https://habrahabr.ru/post/147284/>

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0

6 июля 2012 в 20:33

Разработка → Торговля знает, когда вы ждете ребенка

 Data Mining*

Однажды в магазин Target зашел мужчина и потребовал вызвать менеджера. В своих руках он сжимал огромную кипу купонов магазина, полученных его дочерью.

«Моя дочь получила это по почте! – прокричал он. – Она еще в школу ходит, а вы посылаете ей купоны на детскую одежду и памперсы? Да как вы смеете! Вы хотите таким способом побудить школьниц рожать?»

Менеджер посмотрел на пачку купонов на материнскую одежду, детскую мебель – действительно, они были адресованы дочери рассерженного мужчины. Менеджер принес свои извинения.

Через несколько дней он позвонил мужчине, чтобы еще раз извиниться. По телефону голос отца звучал растерянно. «Знаете, я серьезно поговорил с дочерью, и выяснилось, что в моем доме происходило то, о чем я совершенно не догадывался. Она рождает в августе. Примите мои извинения».

Как Target узнал, что дочь беременна до того, как об этом стало известно ее отцу? Ответ прост – благодаря **системе прогнозирования беременности** (pregnancy prediction system), разработанной аналитиком компании Эндрю Полом. Под катом – обо всем по порядку.

HOME SEARCH

The New York Times Magazine

Magazine

How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012



Antonio Bolfo/Reportage for The New York Times

Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: “If we wanted to figure out if a customer is pregnant, even if she didn’t want us to know, can you do that?”

Pole has a master’s degree in statistics and another in economics, and has

Big Data – Определение #2

3 технологических признака:

1) Гибкое хранилище и возможность делать разнообразные запросы (Mongo, Cassandra, HDFS, Postgres)

2) Надёжное “заведение” данных (Apache Storm, Spark Streaming, Apache Kafka)

Две проблемы распределенных систем

Две проблемы распределенных систем



Две проблемы распределенных систем



Mathias Verraes

@mathiasverraes



Follow

There are only two hard problems in distributed systems: 2. Exactly-once delivery 1. Guaranteed order of messages 2. Exactly-once delivery

В распределенных системах существуют только две серьезные проблемы: 2. Доставить сообщение ровно один раз 1. Гарантировать порядок доставки сообщений 2. Доставить сообщение ровно один раз

Big Data – Определение #2

3 технологических признака:

1) Гибкое хранилище и возможность делать разнообразные запросы (Mongo, Cassandra, HDFS, Postgres)

2) Надёжное “заведение” данных (Apache Storm, Spark Streaming, Apache Kafka)

3) Крутые аналитические инструменты (Hadoop):
query = function(all data)

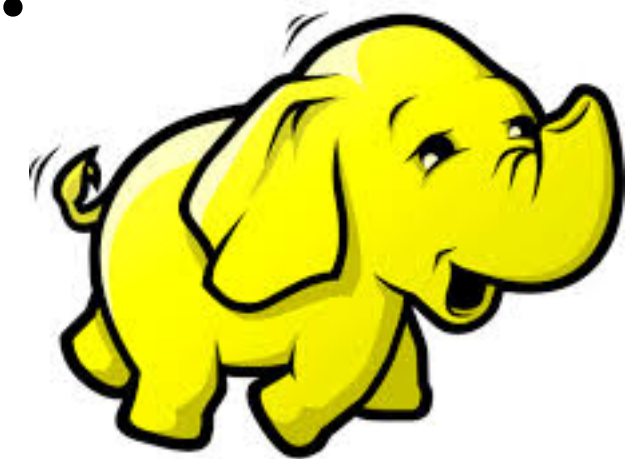
Big Data – Определение #2

3 технологических признака:

1) Гибкое хранилище и возможность делать разнообразные запросы (Mongo, Cassandra, HDFS, Postgres)

2) Надёжное “заведение” данных (Apache Storm, Spark Streaming, Apache Kafka)

3) Крутые аналитические инструменты (Hadoop):
query = function(all data)



Технологии Big Data или Pokemons?



Spoink?



Spoink is a **Pokemon**!



Arvados?



Arvados is a **Big Data**!



Horsea?



Horsea is **Pokemon**!



Seahorse?



Seahorse is a Big Data!



Pangool?



Pangool is a Big Data!



Vulpix?



Vulpix is **Pokemon!**



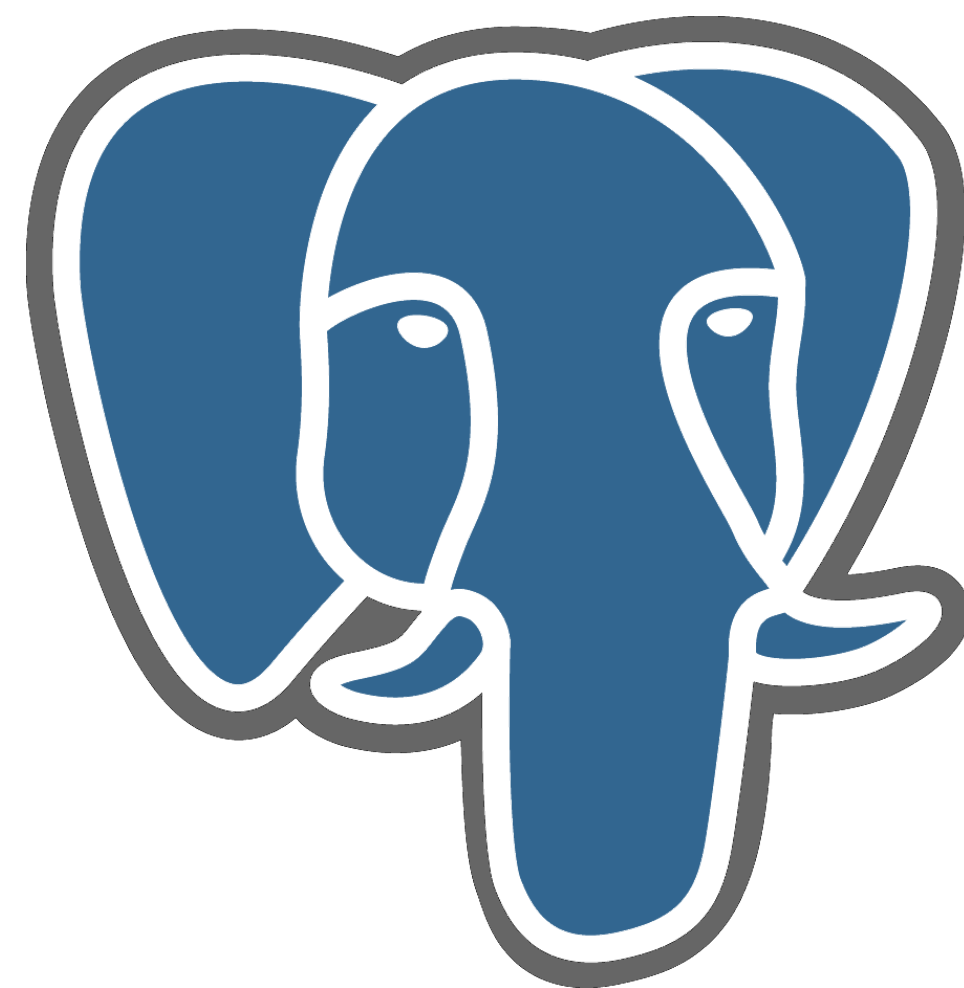
Flink?



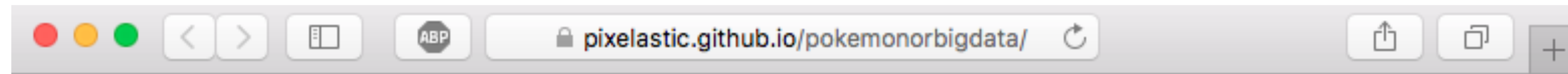
Flink is a Big Data!



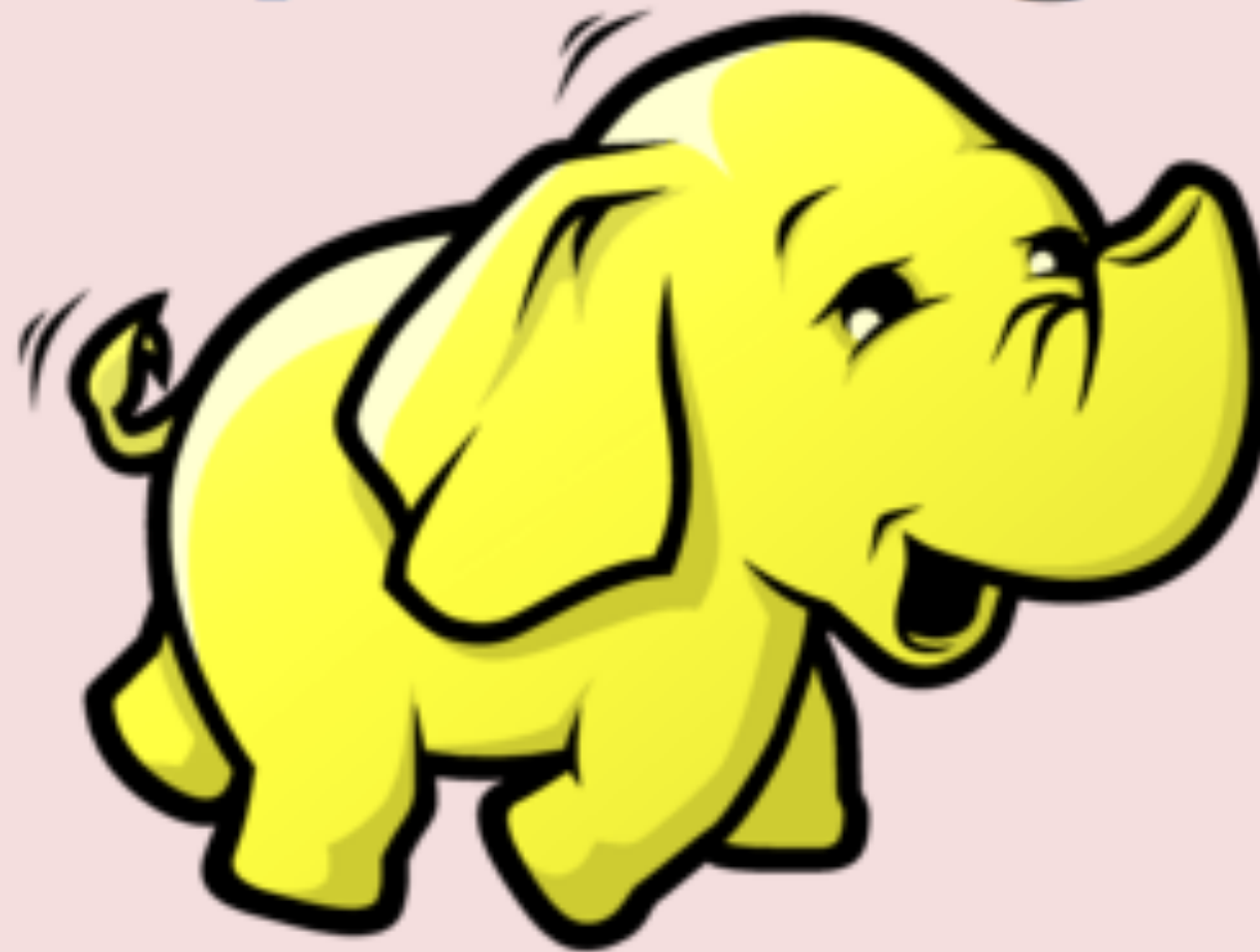
Postgres?



<https://pixelastic.github.io/pokemonorbigdata/>



Hadoop is Big Data!



Hadoop is a distributed system for counting words.

Next question

Made by @pixelastic, inspired by [this google form](#).

Анти-введение в Big Data

Big Data – Определение #3

Big Data – Определение #3

Проектные признаки: 7V

Big Data – Определение #3

Проектные признаки: 7V

1. Volume – Объемы

Big Data – Определение #3

Проектные признаки: 7V

1. Volume – Объемы

2. Variety – Разнообразие

Big Data – Определение #3

Проектные признаки: 7V

- 1. Volume – Объемы
- 2. Variety – Разнообразие
- 3. Veracity – Достоверность

Big Data – Определение #3

Проектные признаки: 7V

- 1. Volume – Объемы
- 2. Variety – Разнообразие
- 3. Veracity – Достоверность
- 4. Velocity – Скорость

Big Data – Определение #3

Проектные признаки: 7V

- 1. Volume – Объемы
- 2. Variety – Разнообразие
- 3. Veracity – Достоверность
- 4. Velocity – Скорость
- 5. Value – Ценность

Big Data – Определение #3

Проектные признаки: 7V

- 1. Volume – Объемы
- 2. Variety – Разнообразие
- 3. Veracity – Достоверность
- 4. Velocity – Скорость
- 5. Value – Ценность
- 6. Volatility – Неустойчивость

Big Data – Определение #3

Проектные признаки: 7V

- 1. Volume – Объемы
- 2. Variety – Разнообразие
- 3. Veracity – Достоверность
- 4. Velocity – Скорость
- 5. Value – Ценность
- 6. Volatility – Неустойчивость
- 7. Visualisation – Визуализация

Big Data – Определение #4

Big Data – Определение #4

Big Data доставляет **головную боль**, в противном случае речь идет о старых добрых ETL, OLAP или CEP

Big Data – Определение #4

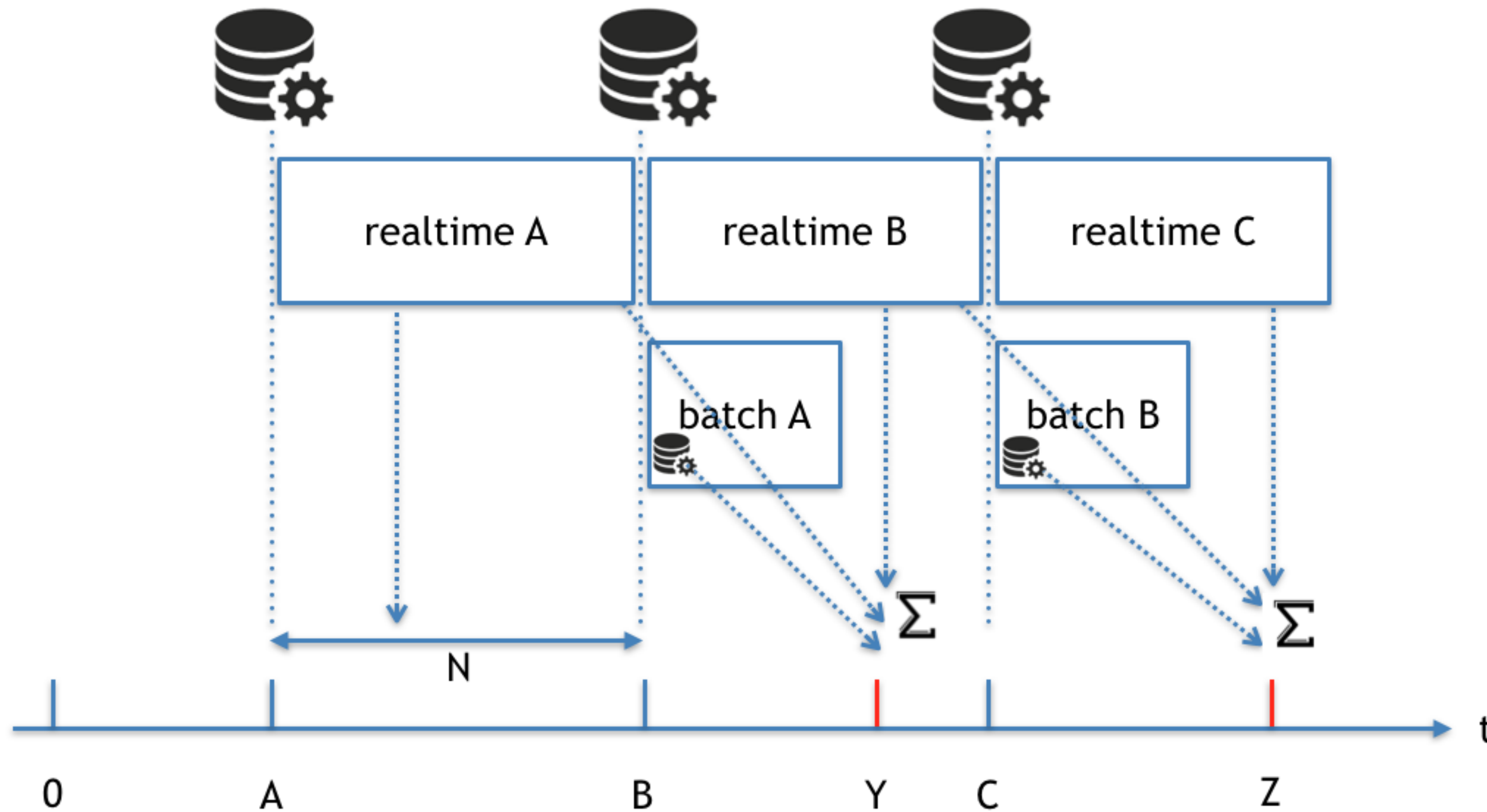
Big Data доставляет **головную боль**, в противном случае речь идет о старых добрых ETL, OLAP или CEP

ETL - Extract Transform Load

OLAP - Online Analytical Processing

CEP - Complex Event Processing

Почему болит голова?



Почему болит голова?



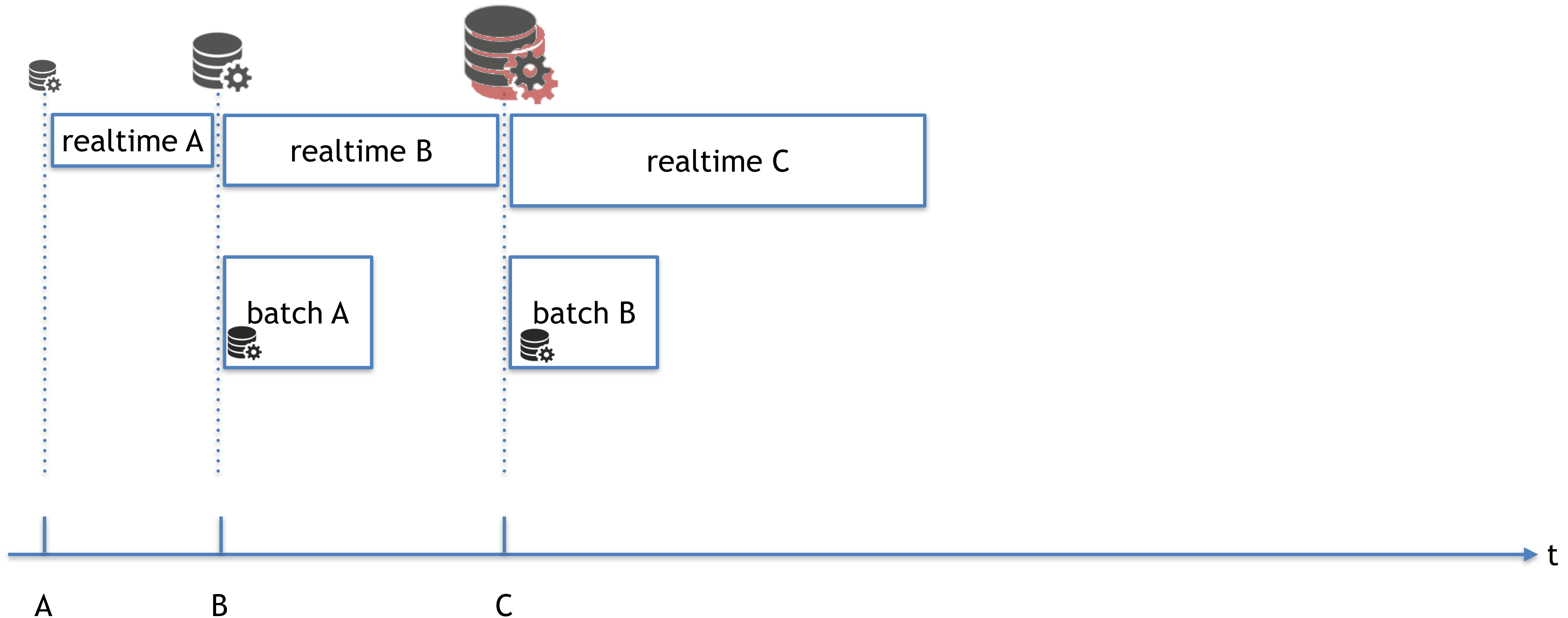
Почему болит голова?



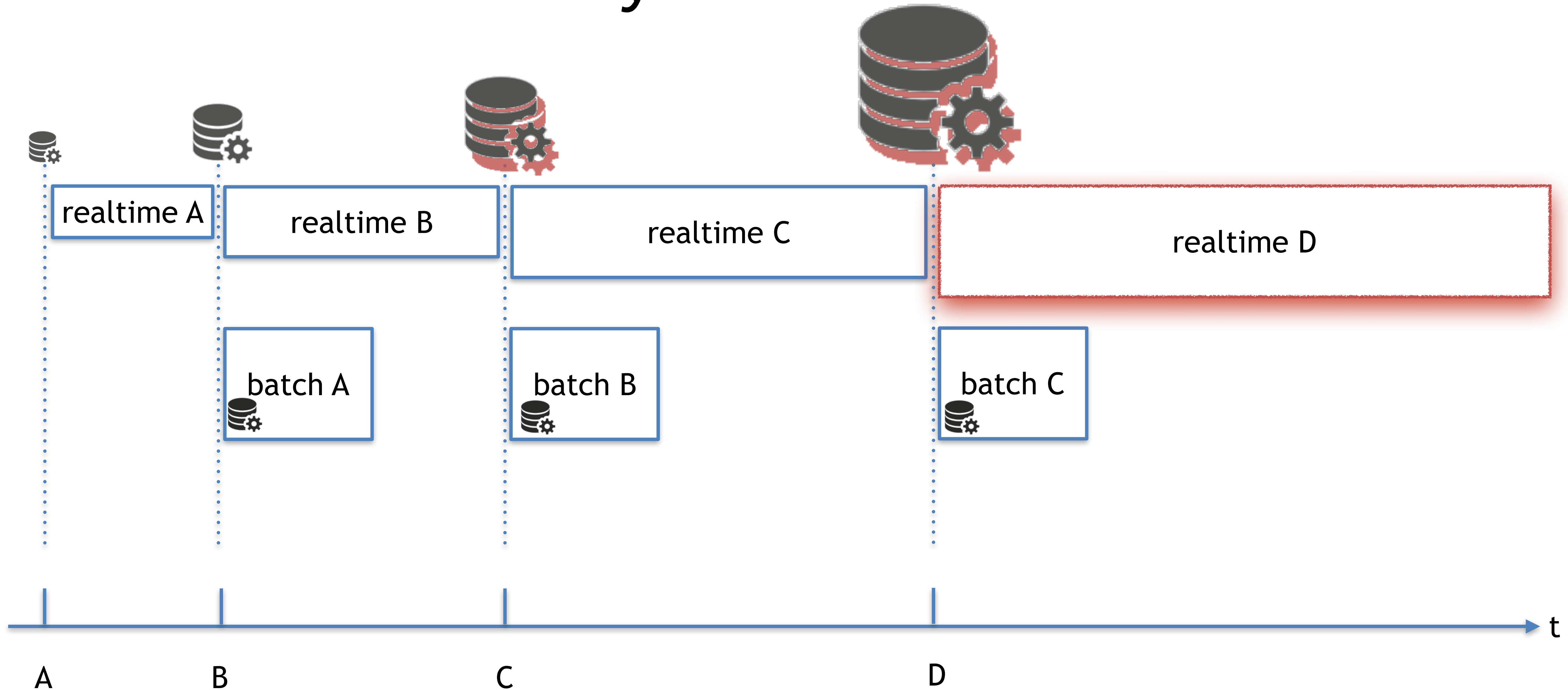
Почему болит голова?



Почему болит голова?



Почему болит голова?

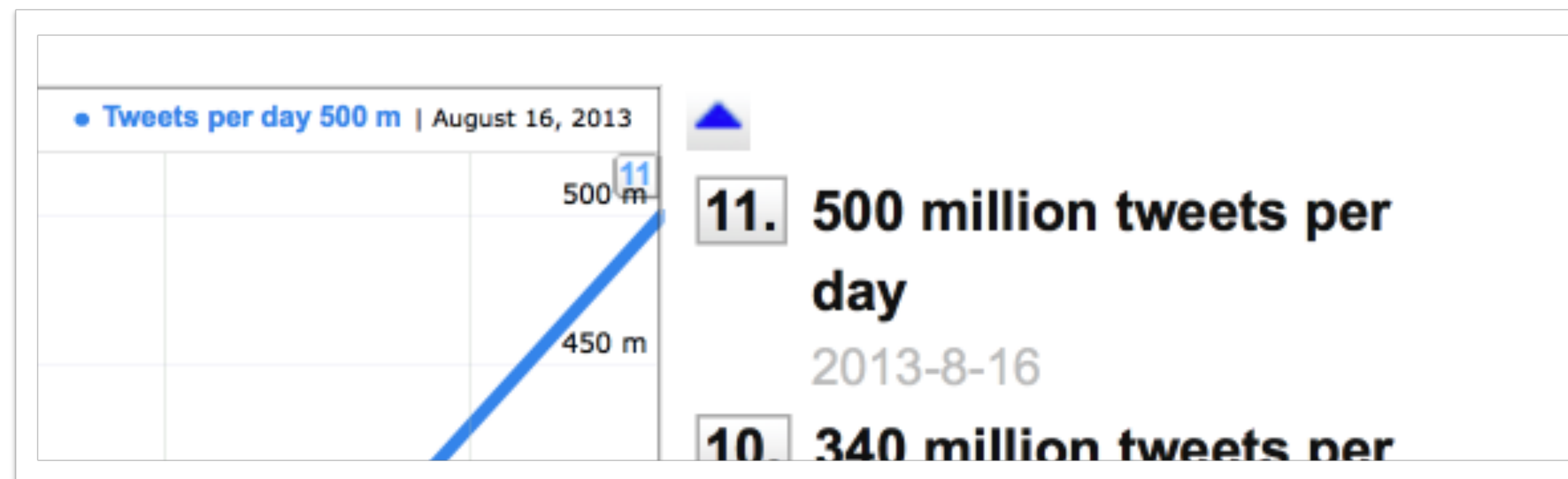


<https://twitter.com/tophashtags> - это Big Data?

 **Top Hashtags** @TopHashtags · 17h

In Australia/6h: 1. #qanda 2. #debate 3. #auspol 4. #4corners 5. #worldmentalhealthday

← ↻ ❤ ...



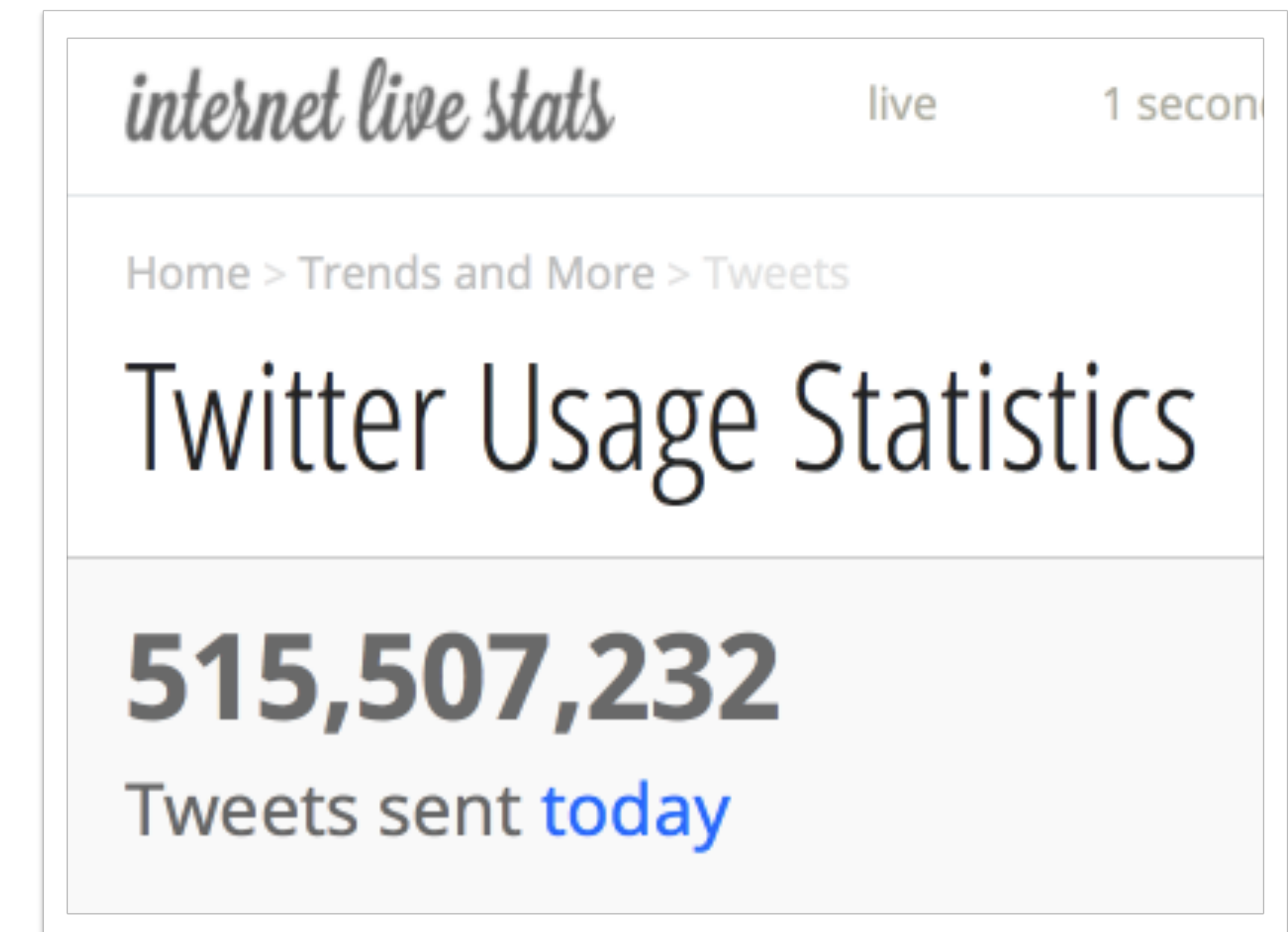
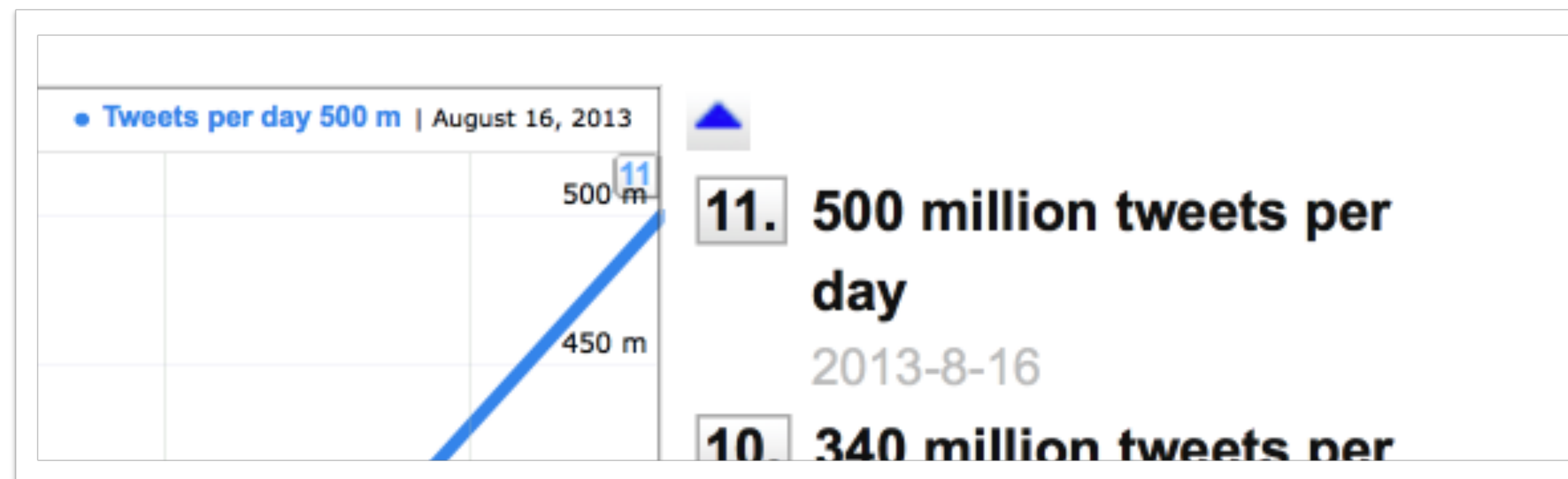
internet live stats live 1 second

Home > Trends and More > Tweets

Twitter Usage Statistics

515,507,232
Tweets sent **today**

<https://twitter.com/tophashtags> - это Big Data?



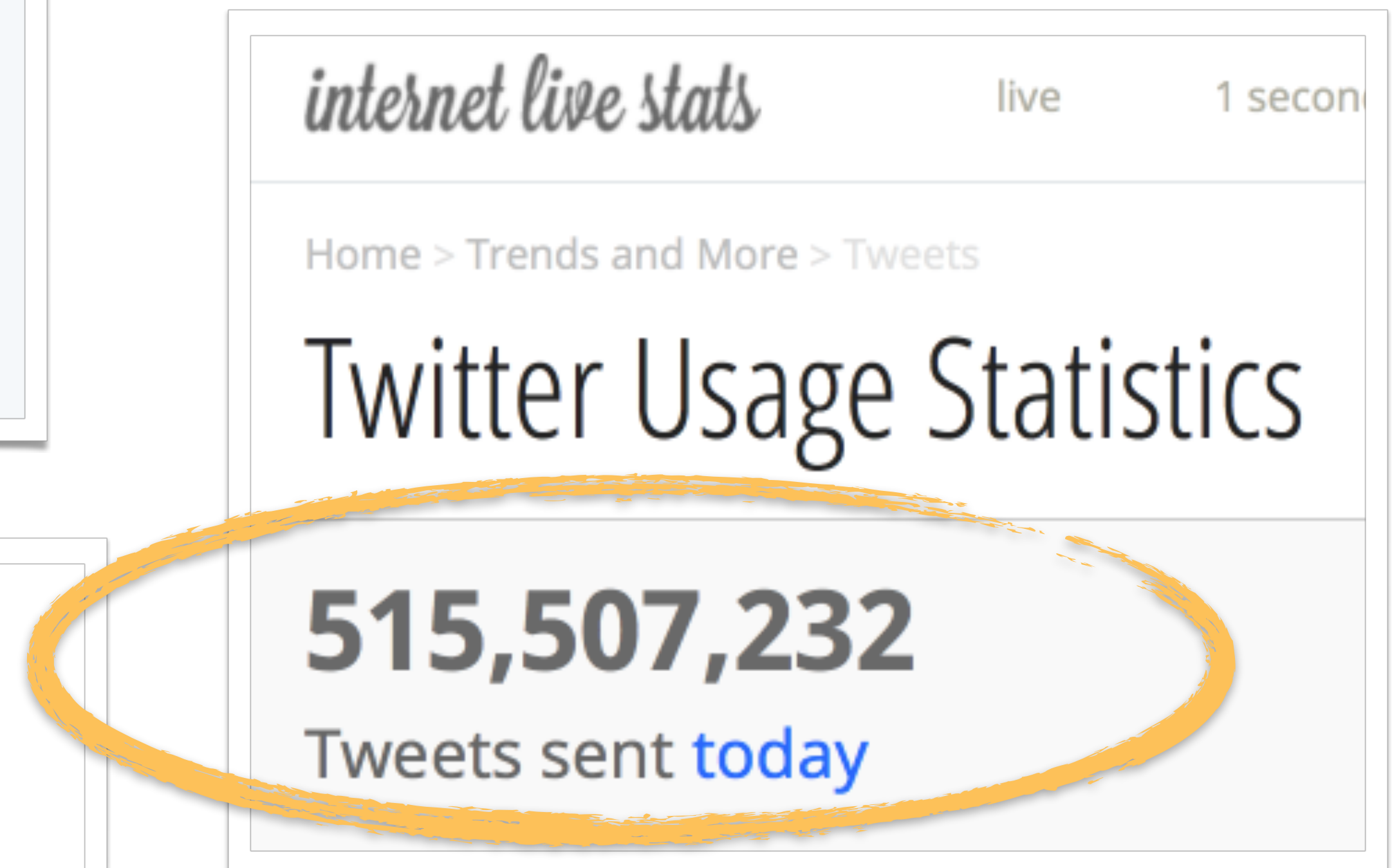
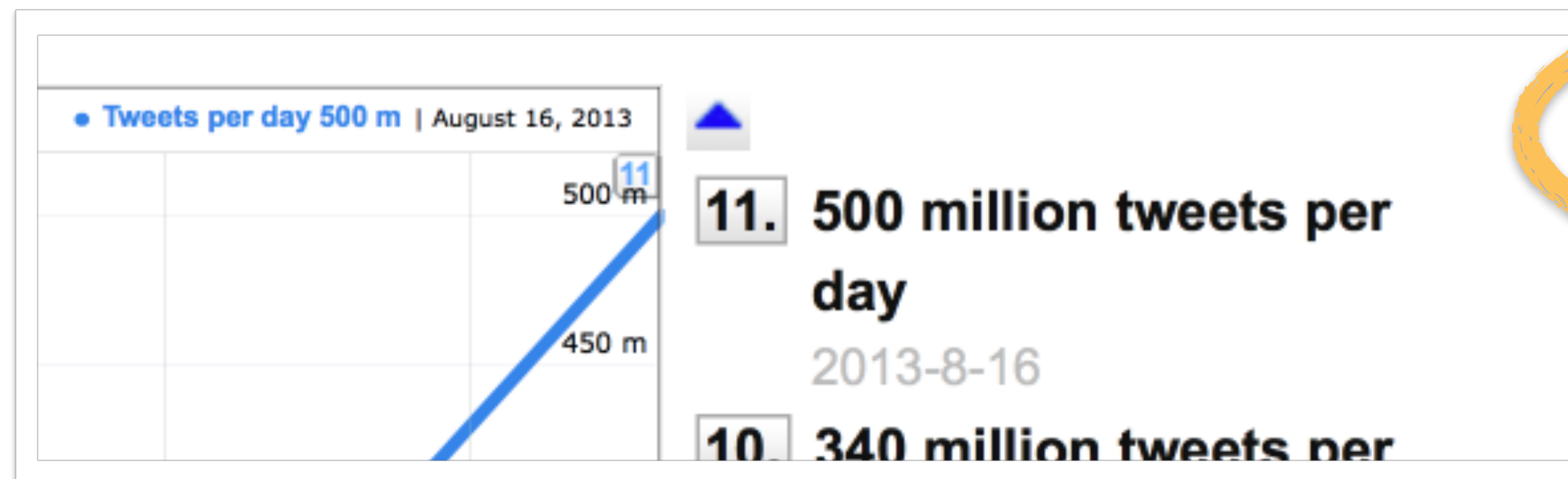
<https://twitter.com/tophashtags> - это Big Data?



A screenshot of a tweet from the account 'Top Hashtags @TopHashtags' posted 17 hours ago. The tweet text is 'In Australia/6h: 1.#qanda 2.#debate 3.#auspol 4.#4corners 5.#worldmentalhealthday'. The entire tweet content is circled in orange. The interface shows standard Twitter interaction icons at the bottom.

Top Hashtags @TopHashtags · 17h

In Australia/6h: 1.#qanda 2.#debate 3.#auspol 4.#4corners 5.#worldmentalhealthday



A screenshot of a 'Twitter Usage Statistics' page. The page title is 'Twitter Usage Statistics'. Below the title, the number '515,507,232' is displayed in large, bold, dark grey font, with the text 'Tweets sent today' in smaller font below it. The entire statistics section is circled in orange.

internet live stats live 1 second

Home > Trends and More > Tweets

Twitter Usage Statistics

515,507,232
Tweets sent today

<https://twitter.com/tophashtags> - это Big Data?



А может Big Data на самом деле
не нужна?



Вы не Яндекс

<https://blog.bradfieldcs.com/you-are-not-google-84912cf44afb>

<https://habrahabr.ru/company/infopulse/blog/330708/>

12 июня в 12:26

Вы — не Google перевод

Проектирование и рефакторинг*, Высокая производительность*, Анализ и проектирование систем*, Google Cloud Platform*,
Блог компании Инфопульс Украина

Мы, программисты, иногда почему-то сходим с ума. Причём по каким-то совершенно нелепым причинам. Нам нравится думать о себе, как о супер-рациональных людях, но когда дело доходит до выбора ключевой технологии нового продукта, мы погружаемся в какое-то безумие. Вдруг оказывается, что кто-то слышал что-то об одной классной вещи, а его коллега читал комментарий о другой на Хабре, а третий человек видел пост в блоге о ещё чём-то похожем... и вот мы уже пребываем в полнейшем ступоре, беспомощно барахтаясь в попытках выбора между совершенно противоположными по своей сути системами, уже и забыв, что мы вообще пытаемся выбрать и почему.

Рациональные люди не принимают решения таким образом. Но именно так программисты часто решают использовать что-то вроде MapReduce.

Вот как [комментировал](#) этот выбор Joe Hellerstein своим студентам (на 54-той минуте):

Дело в том, что в мире сейчас есть где-то 5 компаний, обрабатывающие данные подобных объёмов. Все остальные гоняют все эти данные туда-сюда, добиваясь отказоустойчивости, которая им на самом деле не нужна. Люди страдают гигантоманией и гугломанией где-то с середины 2000-ых годов: «мы сделаем всё так, как делает Google, ведь мы же строим один из крупнейших (в будущем) сервисов по обработке данных в мире!»



Ozan Onay [Follow](#)

Teaching Computer Science at Bradfield

Jun 7 · 7 min read

You Are Not Google

Software engineers go crazy for the most ridiculous things. We like to think that we're hyper-rational, but when we have to choose a technology, we end up in a kind of frenzy—bouncing from one person's Hacker News comment to another's blog post until, in a stupor, we float helplessly toward the brightest light and lay prone in front of it, oblivious to what we were looking for in the first place.

This is not how rational people make decisions, but it *is* how software engineers decide to use MapReduce.

As Joe Hellerstein sideranted to [his undergrad databases class](#) (54 min in):

The thing is there's like 5 companies in the world that run jobs that big. For everybody else... you're doing all this I/O for fault tolerance that you didn't really need. People got kinda Google mania in the 2000s: "we'll do everything the way Google does because we also run the world's largest internet data service" [tilts head sideways and waits for laughter]

Анти-заключение про Big Data

Анти-заключение про Big Data

> Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура
- > Смешные технологии для компонентов лямбда-архитектуры

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура
- > Смешные технологии для компонентов лямбда-архитектуры
- > Много данных из-за избыточности, иммутабельности и репликации

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура
- > Смешные технологии для компонентов лямбда-архитектуры
- > Много данных из-за избыточности, иммутабельности и репликации
- > Технологические признаки: гибкое хранение, надежное заведение данных, крутые аналитические инструменты

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура
- > Смешные технологии для компонентов лямбда-архитектуры
- > Много данных из-за избыточности, иммутабельности и репликации
- > Технологические признаки: гибкое хранение, надежное заведение данных, крутые аналитические инструменты
- > Проектные признаки 7V

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура
- > Смешные технологии для компонентов лямбда-архитектуры
- > Много данных из-за избыточности, иммутабельности и репликации
- > Технологические признаки: гибкое хранение, надежное заведение данных, крутые аналитические инструменты
- > Проектные признаки 7V
- > Головная боль

Анти-заключение про Big Data

- > Постоянная адаптация к смежным задачам и устойчивость к человеческим ошибкам
- > Лямбда-архитектура
- > Смешные технологии для компонентов лямбда-архитектуры
- > Много данных из-за избыточности, иммутабельности и репликации
- > Технологические признаки: гибкое хранение, надежное заведение данных, крутые аналитические инструменты
- > Проектные признаки 7V
- > Головная боль
- > Нужна ли Вам Big Data?



СПАСИБО!

Анти-введение в Big Data



vladimir.krasilschik@gmail.com



@dyer_the

pgday 2017

Санкт-Петербург, Россия

~~Вопросы?~~ Вопросы?





СПАСИБО!

Анти-введение в Big Data

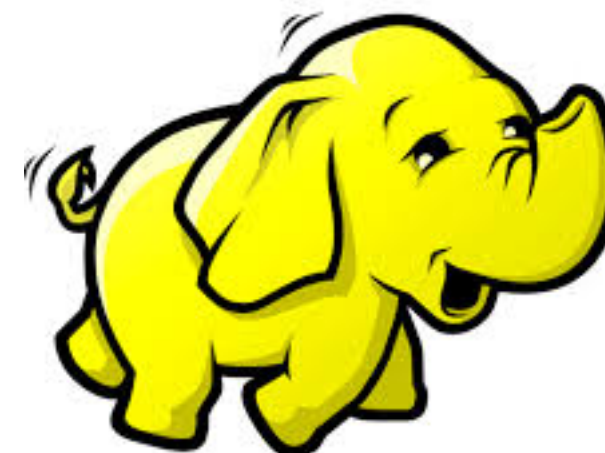


vladimir.krasilschik@gmail.com



@dyer_the
pgday 2017

Санкт-Петербург, Россия



~~Вопросы?~~ Вопросы?





СПАСИБО!

Анти-введение в Big Data

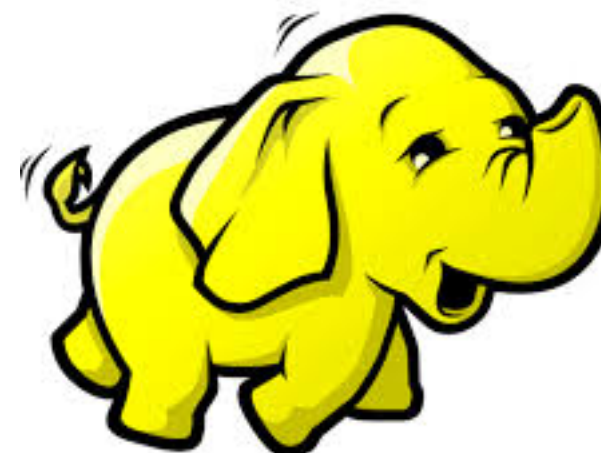


vladimir.krasilschik@gmail.com



@dyer_the
pgday 2017

Санкт-Петербург, Россия



~~Вопросы?~~
Вопросы?





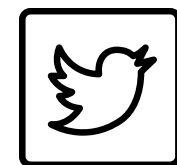
СПАСИБО!



Анти-введение в Big Data



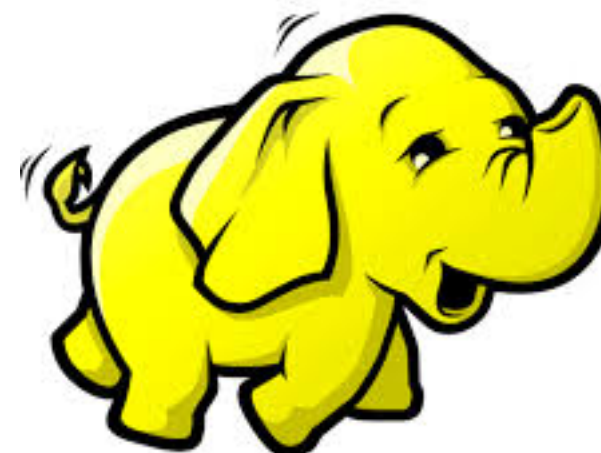
vladimir.krasilschik@gmail.com



@dyer_the

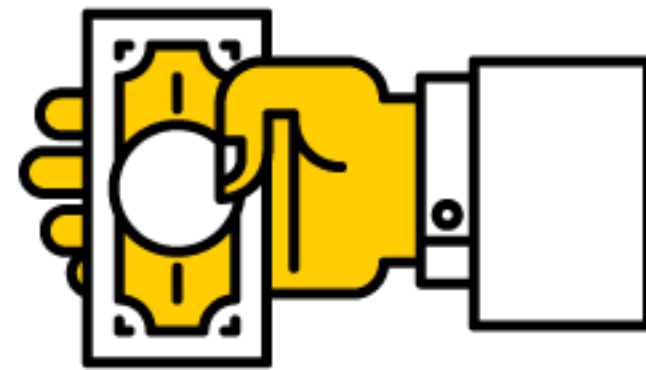
pgday 2017

Санкт-Петербург, Россия



~~Вопросы?~~
Вопросы?





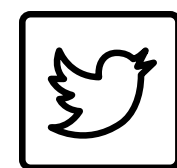
СПАСИБО!



Анти-введение в Big Data



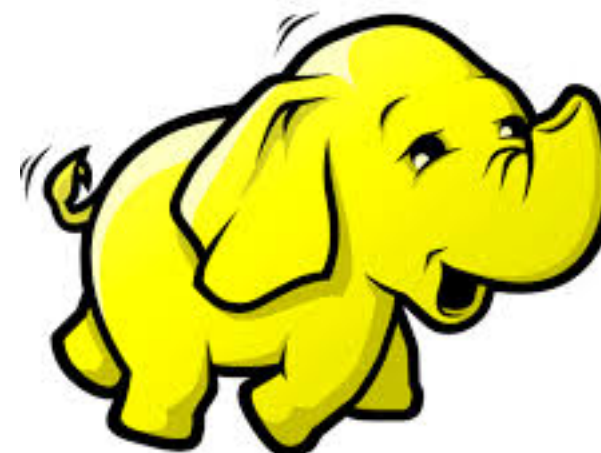
vladimir.krasilschik@gmail.com



@dyer_the

pgday 2017

Санкт-Петербург, Россия



~~Вопросы?~~
Вопросы?

