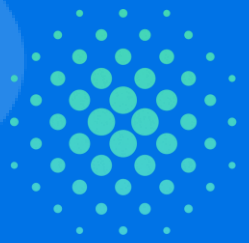
детский мир

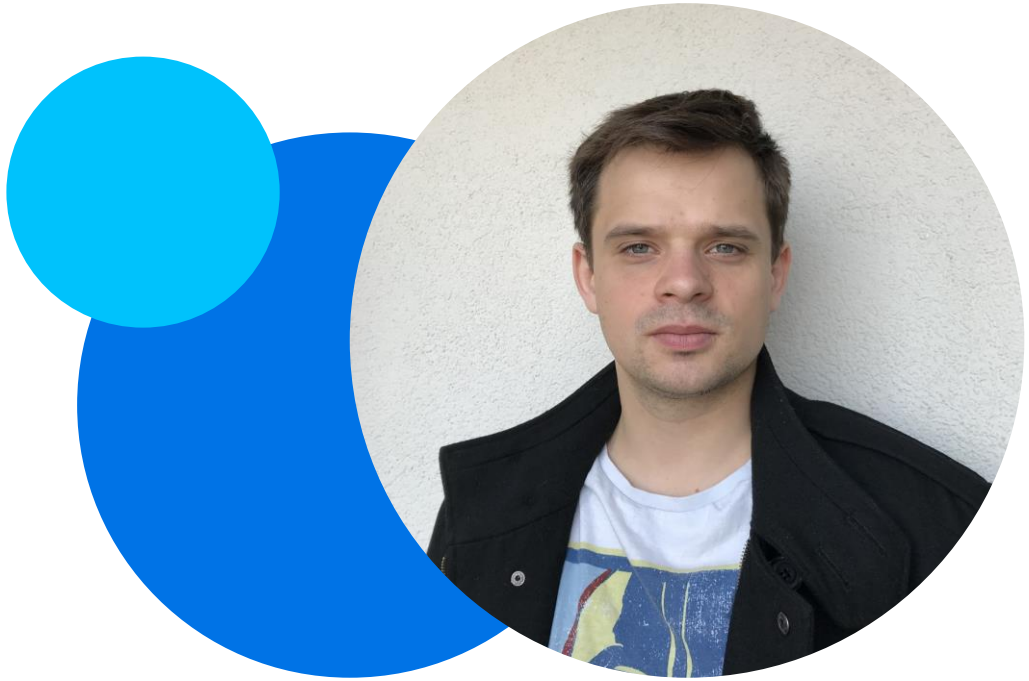
Обработка событий в Snowplow: от сбора до аналитики



SmartData
2024

Седельников Владимир
Бученкова Мария

Докладчики



Седельников Владимир

Ведущий инженер по работе с данными в «Детском мире»
Руководит внедрением кластерного ClickHouse. Участвует в адаптации Snowplow для работы с платформой Hadoop.

Больше сложных задач любит только разводные мосты и щенков овчарок.

Докладчики

Бученкова Мария

Ведущий инженер по работе с данными в «Детском мире». Отвечает за внедрение snowplow на стороне разработки.

В data больше 8 лет.

Прошла путь ML → MLOps → Data engineer и перебрала цвета всех банков.

Структура доклада

Что такое Snowplow и почему мы на нём остановились

Интеграция в инфраструктуру Детского мира

Обработка накопленных событий

Анализ полученных данных. Первые результаты

Структура доклада

Что такое Snowplow и почему мы на нём остановились

Интеграция в инфраструктуру Детского мира

Обработка накопленных событий

Анализ полученных данных. Первые результаты

Структура доклада

Что такое Snowplow и почему мы на нём остановились

Интеграция в инфраструктуру Детского мира

Обработка накопленных событий

Анализ полученных данных. Первые результаты

Структура доклада

Что такое Snowplow и почему мы на нём остановились

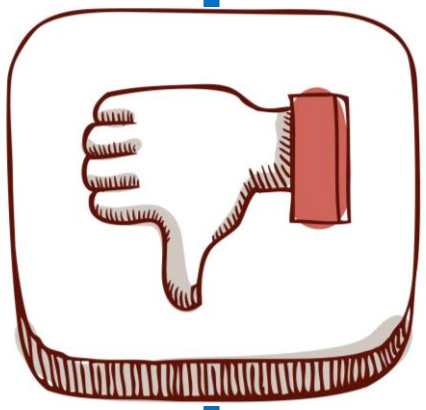
Интеграция в инфраструктуру Детского мира

Обработка накопленных событий

Анализ полученных данных. Первые результаты

Что такое Snowplow и почему мы на нем остановились

Мы долгое время пользовались GA + BQ: почему решили уйти



- Риски
- Стоимость хранения данных
- Стоимость выгрузок

Какие существуют альтернативы GA?

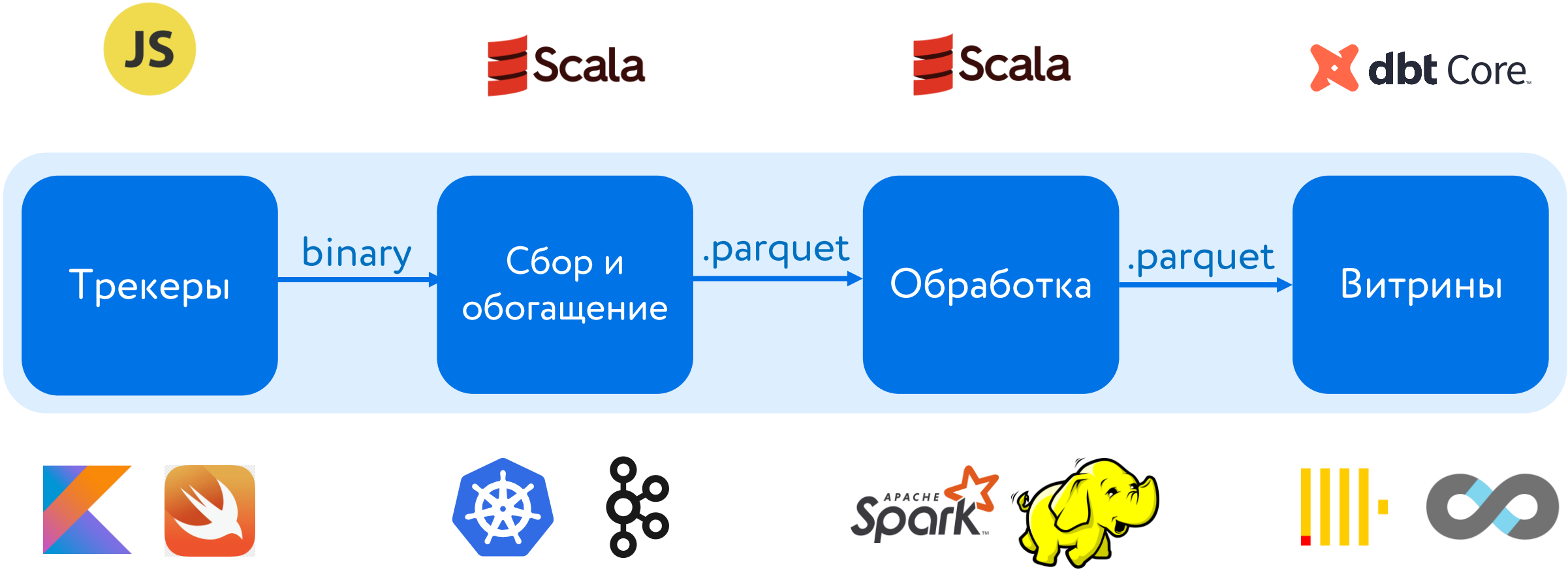
- … Matomo?
- … работает? Не лезь!
- … сделать всё самим с нуля
- … доработать Snowplow под свою архитектуру

Почему именно Snowplow?

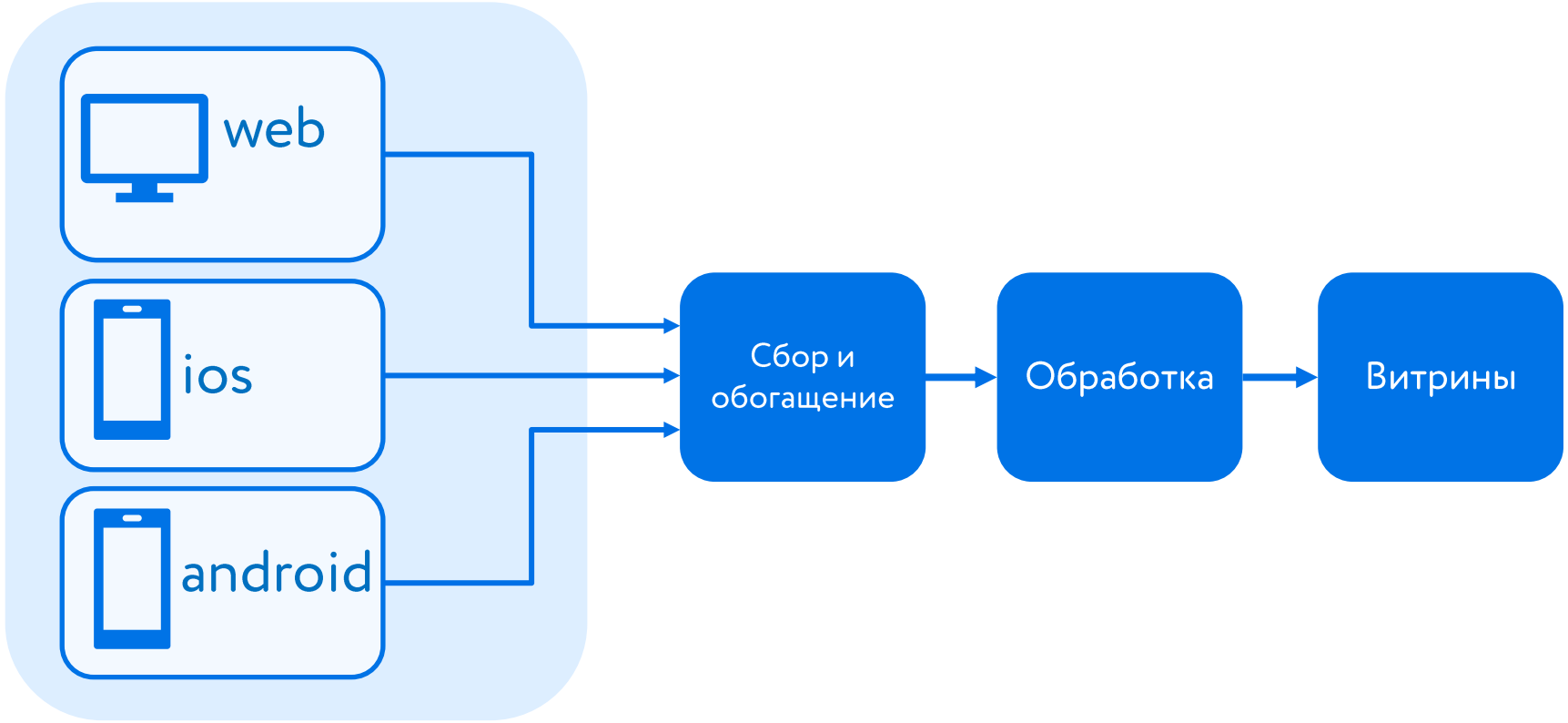
- Импортозамещение
- Все наши данные хранятся у нас
- Использование горизонтально-масштабируемой архитектуры

Интеграция Snowplow в инфраструктуру Детского мира

Схема данных



Трекеры



Android трекеры

Android Tracker SDK поддерживает Android 5 (API level 21+)

SDK есть как на Kotlin, так и на JAVA

```
val tracker = Snowplow.createTracker(
    applicationContext, // Android context (LocalContext.current in Compose apps)
    "appTracker", // namespace
    "https://snowplow-collector-url.com" // Event collector URL
)
```

```
val event = ScreenView("screen_name")
tracker.track(event)
```

```
Snowplow.defaultTracker?.track(event)
```

iOS трекеры

Snowplow iOS Tracker SDK поддерживает iOS 11.0+, macOS 10.13+, tvOS 12.0+, watchOS 6.0+, и visionOS 1.0+

SDK есть как для Swift, так и для Objective-C

```
let tracker = Snowplow.createTracker(namespace: "appTracker", endpoint: "https://snowplow-collector-url.com")
```

```
let event = ScreenView(name: "screen_name")
tracker.track(event)
```

```
Snowplow.defaultTracker()?.track(event)
```

WEB трекеры

Snowplow предоставляет WEB трекеры на JavaScript

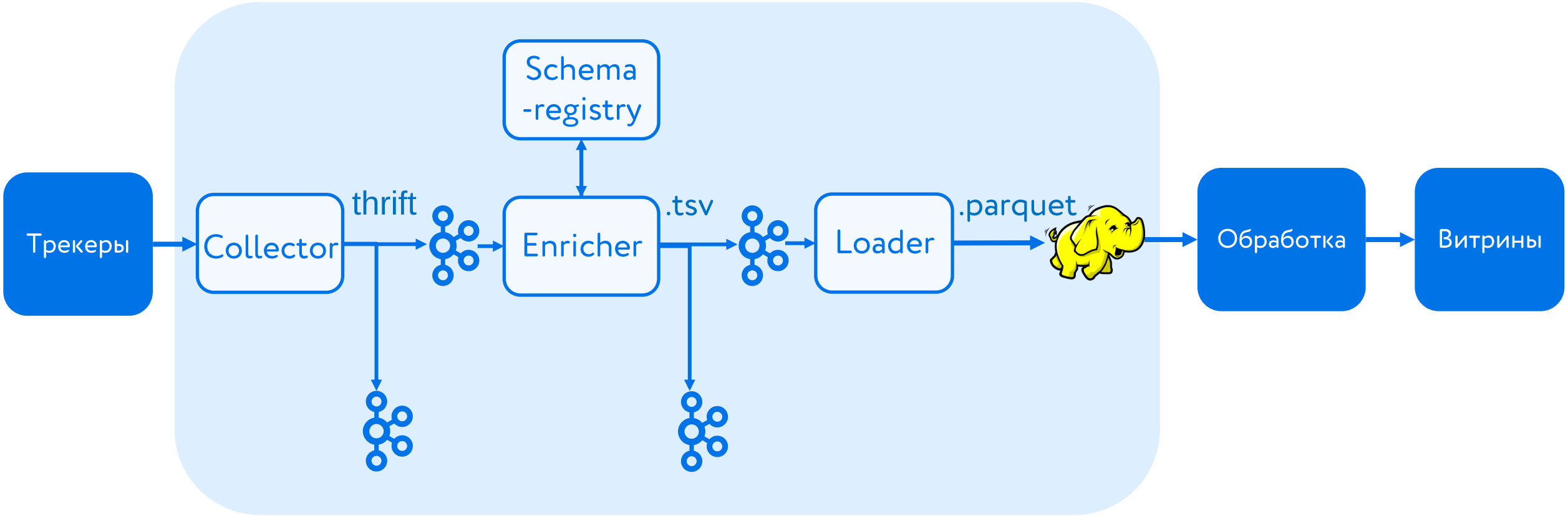
Пример подключения JavaScript tracker

```
;(function(p,l,o,w,i,n,g){if(!p[i]){p.GlobalSnowplowNamespace=p.GlobalSnowplowNamespace||[];
p.GlobalSnowplowNamespace.push(i);p[i]=function(){(p[i].q=p[i].q||[]).push(arguments)
};p[i].q=p[i].q||[];n=l.createElement(o);g=l.getElementsByTagName(o)[0];n.async=1;
n.src=w;g.parentNode.insertBefore(n,g)}}(window,document,"script","{{URL to sp.js}}","snowplow"));
```

```
window.snowplow('newTracker', 'sp1', '{{collector_url}}', {
  appId: 'my-app-id'
})
```

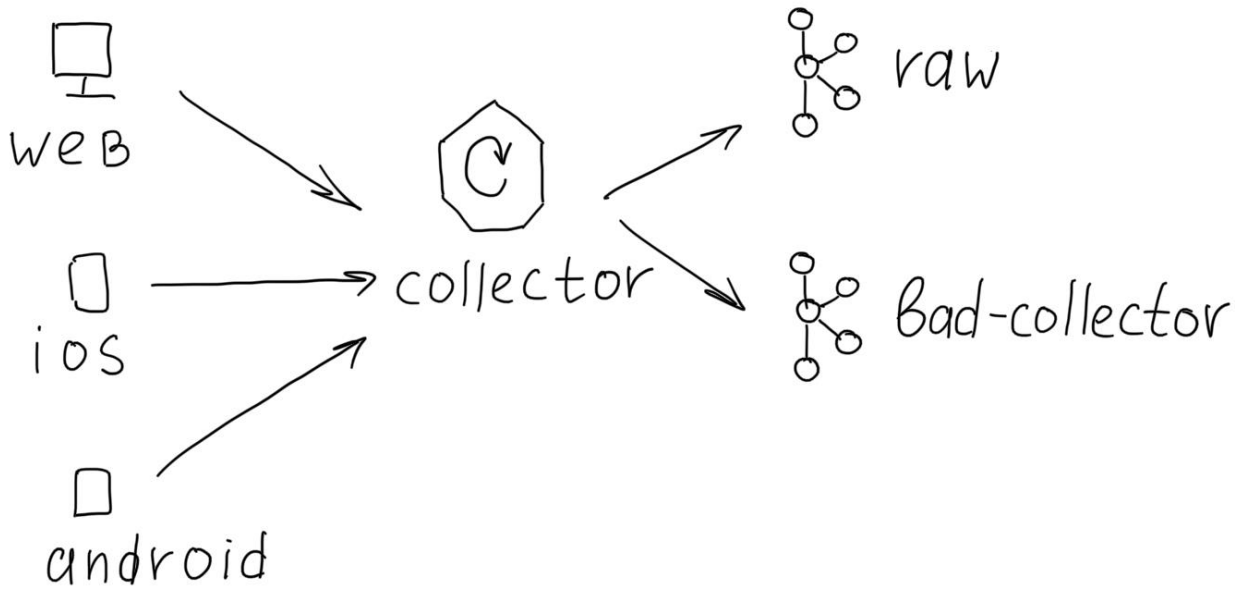
```
window.snowplow('trackPageView');
```

Сбор и обогащение



Collector

- Сервис: Snowplow
- Задача: максимально быстро принять запрос
- Вход: данные с клиентских устройств
- Выход: топики kafka (Thrift - формат сообщений)



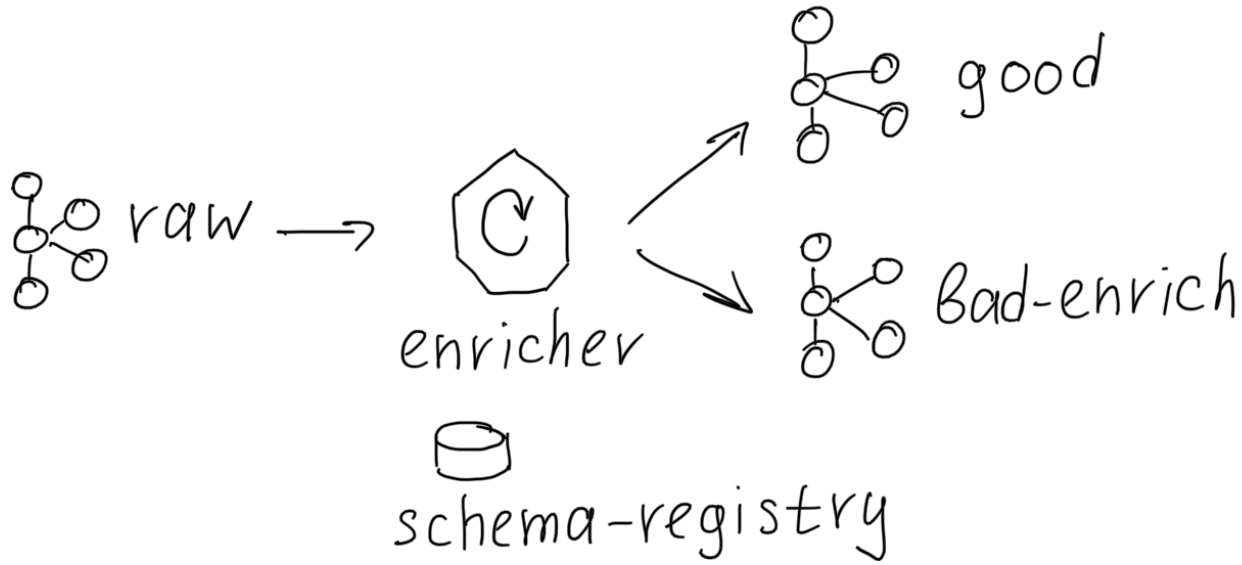
Collector

	Key	Timestamp	Partition	Offset	Headers	Schema	
☐	176.15.210.224	3 days ago	4	1237085399	0		

d 176.15.210.224
UTF-8 ssc-2.9.2-kafka ,2DetmirStore/72081 CFNetwork/1410.0.3 Darwin/22.6.0 @#/com.snowplowanalytics.snowplow/tp2 T {"schema":"iglu:com.snowplowanalytics.snowplow\/payload_data\/jsonschema\/1-0-4","data":
[{"vp":"1170x2532","eid":"2F0270D6-7741-4FA7-A670-653D51E80989","p":"mob","cx":"eyJzY2hlbWEiOiJpZ2x1OmNvbS5zbm93cGxvd2FuYWx5dGljcy5zbm93cGxvd1wvY29udGV4dHNcL2pzb25zY2hlbWFcLzEtMC0xIiwiZGF0YSI6W3sic2NoZW1hIjoiaWdsdTpjb20uc25vd3Bsb3dhbmFseXRpY3Muc25vd3Bsb3dcL21vYmlsZV9jb250ZXh0XC9qc29uc2NoZW1hXC8xLTAtMiIsImRhdGEiOnsic5c2ljYWxNZW1vcnkiOjU5MjY5Nzc1MzYsImFwcGxlSWRmdiI6IkRBQzFCQThELTY2Q0QtNDlERi04NEEwLUQ1N0EyMzRDMjhGMCIsImJhdHRlcnlfdGF0ZSI6ImNoYXJnaW5nIiwib3NWZXJzaW9uIjoiMTYuNi4xIiwiYXBwQXZhaWxhYmxlTWVtb3J5IjoyNzUxNTkwMjcyLCJuZXR3b3JrVHlwZSI6IndpZmkiLCJvc1R5cGUiOiJpb3MiLCJkZXZpY2VNb2RlbCI6ImlQaG9uZTE0LDIiLCJ0b3RhbFN0b3JhZ2UiOjI1NTg2NzU3NDQsImlkZW50aWZpZXJGb3JWZW5kb3IiOiJBNjJFMjhGMC1CMjhGLTQ2QzAtQjQ0Ri1BMjhGMEIyOEY2RTAiLCJ0aW1lem9uZSI6IkFtZXJpY2FcL0xvc19BbmdlbGVzIiwiYXBwQnVpbGRJZCI6IjIwRzgxIiwiYXBwSW5zdGFsbFRpbWUiOjE2NzU0NjQ0MDAuMDAwLCJ0aW1lem9uZU9mZnNldCI6LTQ4MDAsImFwcFZlcnNpb24iOiIyLjE0LjAi

Энричер

- Сервис: Snowplow
- Разгребает то, что пришло с коллектора
- Проверяет соответствие событий json схеме
- Обогащает пользовательскими данными
- Вход: сообщения с коллектора в Thrift формате
- Выход: отдельные events в tsv

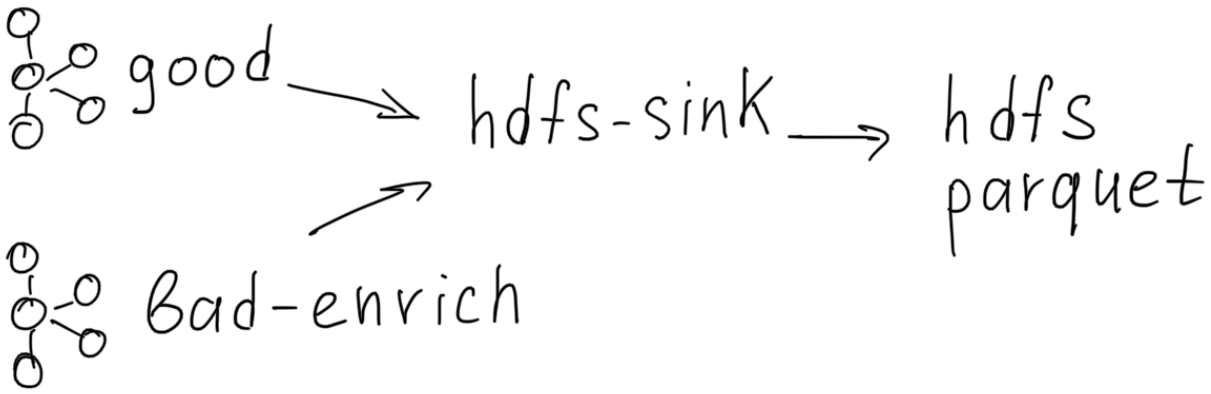


Энричер

```
detmir_and_ru mob 2024-08-15 08:09:34.186 2024-08-15 07:02:27.000 2024-08-15 07:02:23.118 unstruct
  3706888b-2995-45a1-acaa-8ef67bca83f4    detmir_and_ru andr-5.4.2 10.2.6 ssc-2.9.2-kafka snowplow
  -stream-enrich-3.8.0    5.141.192.177       f600fe54-970c-4f18-adc0-fde57dd735b7
                                                {"schema":"iglu:com.snowplowanalytics
  .snowplow/contexts/jsonschema/1-0-1","data":[{"schema":"iglu:ru.expf/user_props/jsonschema/2-0-0"
  ,"data":{"user_id":"3986856","user_phone":"          ","experiment_name":""
  ,"experiment_variant":"","user_basket_id":"0e5017c122cba5baf3ce2956f29fcf95.123","bonus_card_id"
  :"2221126378931382","basket_fullness":"full"}},{"schema":"iglu:com.snowplowanalytics.snowplow
```

Loader

- Не было стандартного загрузчика в hdfs
- Наш загрузчик - spark structured-streaming job
- Вход: топики kafka good + bad-enrich
- Выход: hdfs

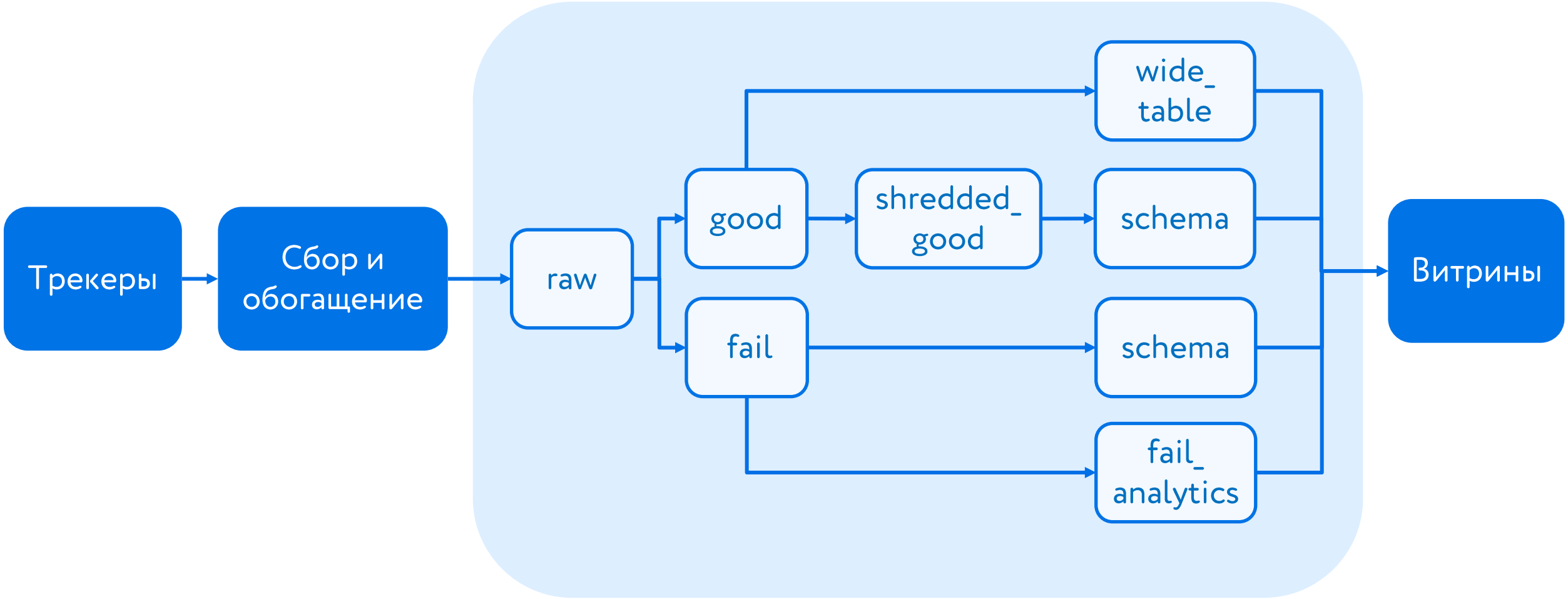


Loader

```
-rw-r--r--   3 bigdata_prod supergroup  240335672 2024-08-17 00:06 /ods/snowplow/01_raw_data/date=2024-08-16/part-00003-f1698c32-2356-4ac0-9d54-58d8369f3920.c000.snappy.parquet
-rw-r--r--   3 bigdata_prod supergroup  900594653 2024-08-17 00:07 /ods/snowplow/01_raw_data/date=2024-08-16/part-00003-f1dcc8f6-26fe-47f8-9270-5bfbd82ff4e2.c000.snappy.parquet
-rw-r--r--   3 bigdata_prod supergroup  878703407 2024-08-17 00:08 /ods/snowplow/01_raw_data/date=2024-08-16/part-00003-f23add8a-0b12-4906-a135-fb5f5a97bc14.c000.snappy.parquet
-rw-r--r--   3 bigdata_prod supergroup 2097771855 2024-08-17 00:07 /ods/snowplow/01_raw_data/date=2024-08-16/part-00003-f243b58c-bd5e-4f1c-8db7-628351ad4958.c000.snappy.parquet
-rw-r--r--   3 bigdata_prod supergroup 1410652496 2024-08-17 00:05 /ods/snowplow/01_raw_data/date=2024-08-16/part-00003-f3578881-a217-4e32-96a5-216b3bca747d.c000.snappy.parquet
```

key	event	topic	partition	offset	timestamp
91.223.89.111	ru.detmir.app\tmo...	snowplow.prod.sta...	7	1805656361	2024-08-16 01:32:...
128.204.79.240	detmir_and_ru\tmo...	snowplow.prod.sta...	7	1805656362	2024-08-16 01:32:...
128.204.79.240	detmir_and_ru\tmo...	snowplow.prod.sta...	7	1805656363	2024-08-16 01:32:...
128.204.79.240	detmir_and_ru\tmo...	snowplow.prod.sta...	7	1805656364	2024-08-16 01:32:...
217.66.159.123	detmir_and_ru\tmo...	snowplow.prod.sta...	7	1805656365	2024-08-16 01:32:...

Обработка накопленных событий



Фильтрация + парсинг tsv

- Good - raw (topic good)

```
+------------+--------------------+--------------------+--------------------+
|      app_id|            contexts|    derived_contexts|      unstruct_event|
+------------+--------------------+--------------------+--------------------+
|detmir_and_ru|{"schema":"iglu:c...|{"schema":"iglu:c...|{"schema":"iglu:c...|
|detmir_and_ru|{"schema":"iglu:c...|{"schema":"iglu:c...|{"schema":"iglu:c...|
|detmir_and_ru|{"schema":"iglu:c...|{"schema":"iglu:c...|{"schema":"iglu:c...|
|detmir_and_ru|{"schema":"iglu:c...|{"schema":"iglu:c...|{"schema":"iglu:c...|
|detmir_and_ru|{"schema":"iglu:c...|{"schema":"iglu:c...|{"schema":"iglu:c...|
+------------+--------------------+--------------------+--------------------+
```

- Fail_analytics - raw (topic bad-enrich)

```
+--------------------+--------------------+--------------------+--------------------+
|               error|   failure_timestamp|             tracker|              schema|
+--------------------+--------------------+--------------------+--------------------+
|{"error":"Validat...|2024-09-04 12:13:...|andr-3.1.1 8.5.2....|iglu:com.snowplow...|
|{"error":"Resolut...|2024-09-04 12:13:...|           js-3.12.0|iglu:com.acme_com...|
|{"error":"Validat...|2024-09-04 12:15:...|         andr-3.1.1|iglu:com.snowplow...|
|{"error":"Resolut...|2024-09-04 12:13:...|           js-3.12.0|iglu:com.acme_com...|
|{"error":"Validat...|2024-09-04 12:13:...|ios-4.0.0 2.72.1....|iglu:com.snowplow...|
+--------------------+--------------------+--------------------+--------------------+
```

Маленькая проблема

Good **МНОГО** весит в hdfs

Выделение плоской структуры

- Вход:
  - простые поля (atomic)
  - поля со вложенной структурой (contexts, derived_contexts, unstruct_event)
- Выход: у всех колонок плоская структура

Выделение плоской структуры

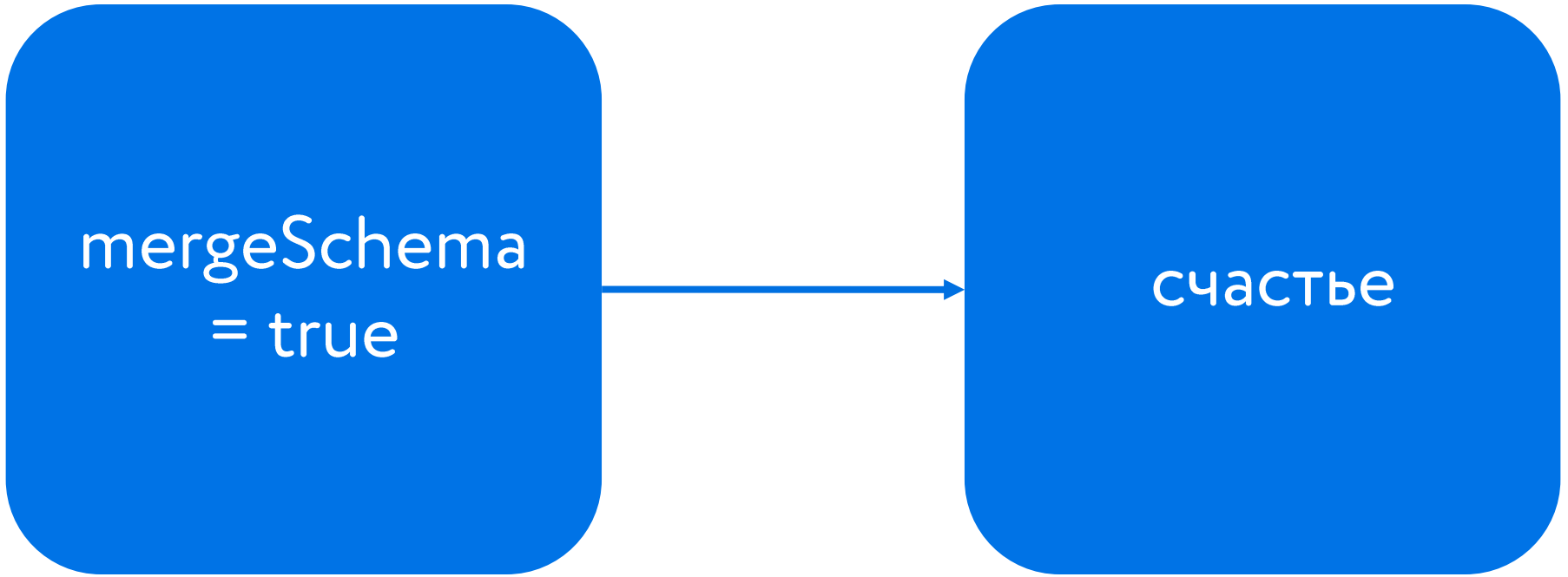
Пример derived_contexts на входе:

{«schema»:
«iglu:com.snowplowanalytics.snowplow/contexts/jsonschema/1-0-1»,
«data»: [{«schema»: «iglu:nl.basjes/yauaa_context/jsonschema/1-0-4»,
«data»:{«deviceBrand»: «Unknown» ,
«deviceName»: «Android Mobile»,
«operatingSystemVersionMajor»: «10»,
«layoutEngineNameVersion»: «AppleWebKit ??»

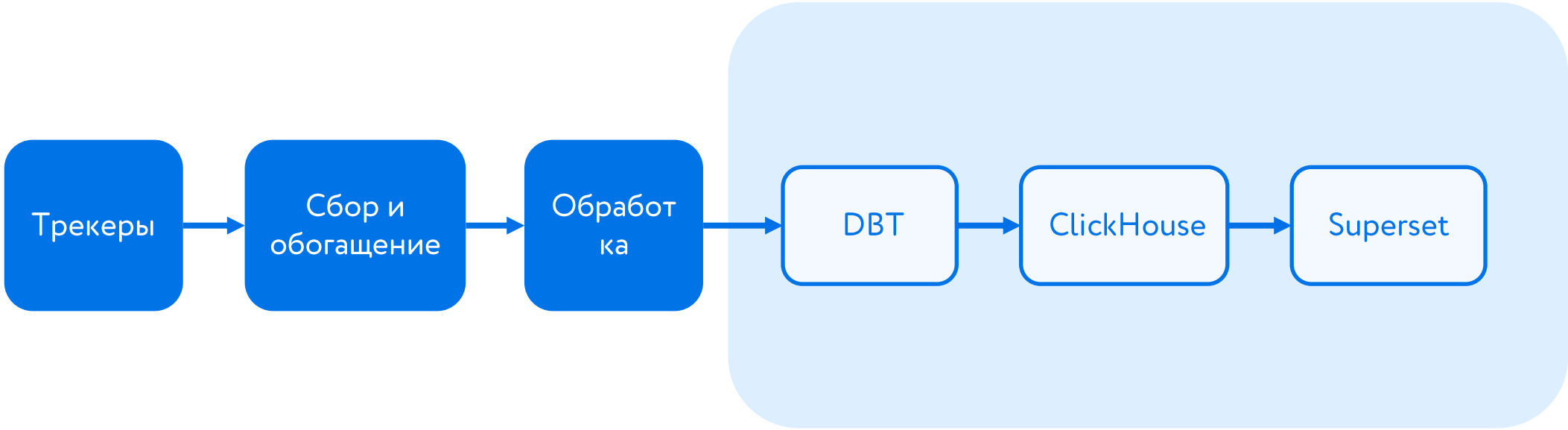
```
root
 |-- unstruct_event_com_snowplowanalytics_snowplow_link_click_1_0_1: struct (nullable = true)
 |    |-- elementId: string (nullable = true)
 |    |-- elementClasses: array (nullable = true)
 |    |    |-- element: string (containsNull = true)
 |    |-- elementTarget: string (nullable = true)
 |    |-- targetUrl: string (nullable = true)
 |    |-- elementContent: string (nullable = true)
 |-- contexts_org_ietf_http_cookie_1_0_0: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- name: string (nullable = true)
 |    |    |-- value: string (nullable = true)
```

Создание супер-схемы

- В ежедневной партиции - разное число колонок
- Пользователи хотят пользоваться



Построение витрин



Анализ событий. Нюансы с тригерами

Сверка данных App

ios. Есть ли на платформе?	android. Есть ли на платформе	ios		android	
			Корректно?		Корректно?
click_promo	click_promo	-9.25%	⚠	-0.21%	✓
select_delivery_method	select_delivery_method	-0.95%	✓	7.49%	⚠
view_checkout	view_checkout	0.01%	✓	-3.51%	⚠
view_delivery_method	N/A	1.2%	✓	-	⚠

Анализ событий. Нюансы с тригерами

Сверка данных Web

Event (SP)	eventCategory (GA)	eventAction (GA)	События разницы %	Пользователи разница %	События сошлись	Пользователи сошлись
add_to_cart	ecommerce	add	2,46%	0,85%	✓	✓
search_click_suggest	search	click_suggest	-82,98	1,58%	⚠	✓
view_cart	ecommerce	checkout	-0,01%	9,16%	✓	⚠
change_bonus_region	account	change_bonus_region	3412,66%	0,78%	⚠	✓

Пакеты dbt snowplow

Метрики, которые рассчитываются «из коробки»:

- Screen views
- Sessions
- Users
- App errors
- Page views
- Consent

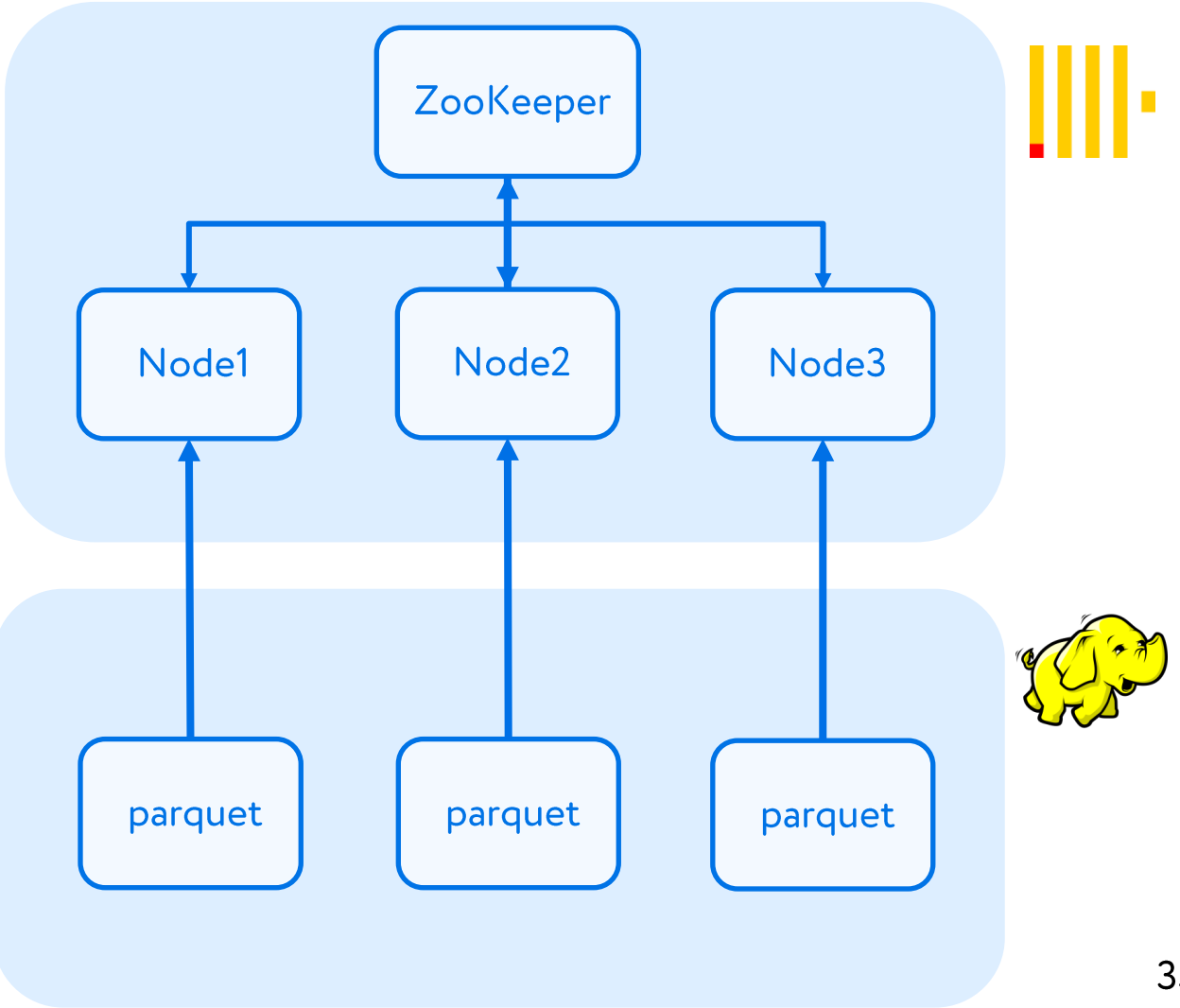
The supported data warehouses per version can be seen below:

Snowplow Unified Digital | Snowplow Media Player | Snowplow Normalize | Snowplow E-commerce | Snowplow Attribution

snowplow-unified version	dbt versions	BigQuery	Databricks	Redshift	Snowflake	Postgres
0.4.5	>=1.6.0 to <2.0.0	✅	✅	✅	✅	✅

Файловый загрузчик в ClickHouse

- Создание/Обновление таблицы
- Распределение файлов между рабочими нодами
- Вставка .parquet файлов в таблицу



Результаты работы

- Количество событий непрерывно растет.
  Сейчас это от 500 млн – до 1.1 млрд
- 45 – 110 Гб в сутки – отсортированные .parquet файлы,
  сжатые кодеком zstd
- Обработка данных за сутки - около 5 часов.
  100 ядер, 32 GB оперативной памяти
- Сейчас у нас развернуто в Kubernetes 10 колекторов и 50 энричеров

Выводы

- Изменилась лицензия с 2024.01.08
- Нужна своя инфраструктура и ресурсы
- Open Source + Scala – будут проблемы, и они будут ваши
- Отсутствие классных UI и встроенных интеграций

- Кастомизируйте
- Масштабируйте
- Минимизируйте задержку
- Большим компаниям стоит присмотреться
- Хранить данные у себя – бесценно

Спасибо!