Manychat

Моделирование потока событий

Об авторе

Manychat

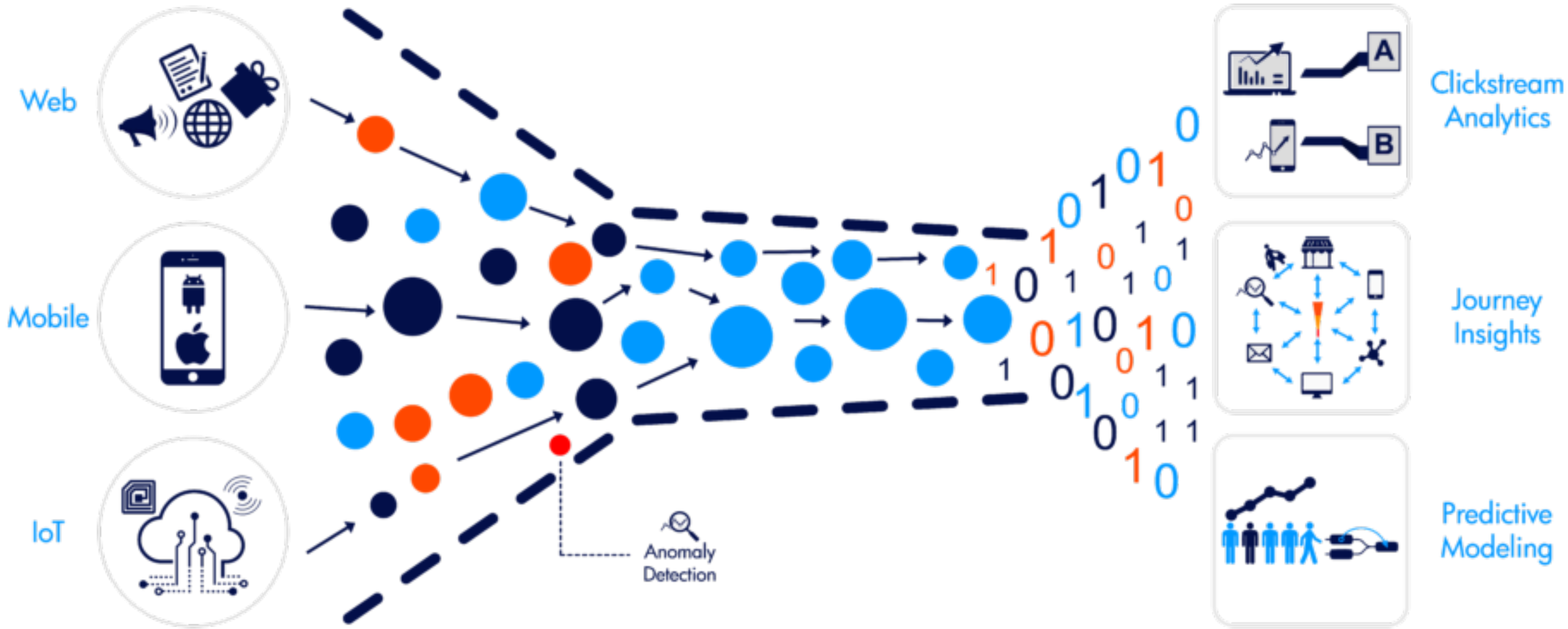
Nikolai Golov

……..-2013	teradata.	Data Vault	<10Tb
2013-2019	VERTICA	Anchor Modeling	11->500Tb
2019-now	snowflake	Anchor Modeling+	3->600Tb

ManyChat

Clickstream - поток событий

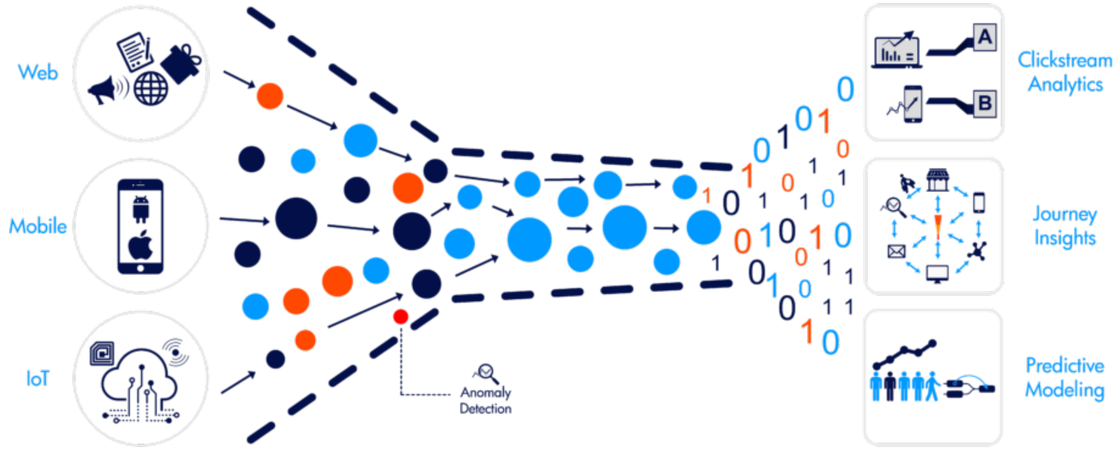
Manychat



Характеристики clickstream:

Manychat

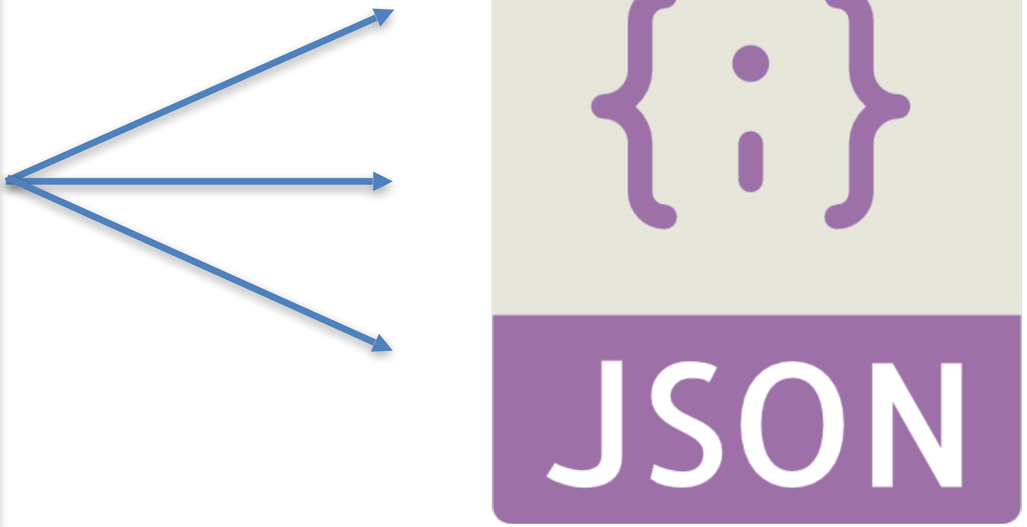
1. Сотни (тысячи) типов событий
2. Схема каждого типа события меняется
3. Аналитики хотят одну таблицу… быструю и простую



Manychat

Semi-Structured clickstream (JSONs)

Event:
-cookie
-URL
-Referer
-Event_type
-Event_dt
-IP
-User_Agent
-.....



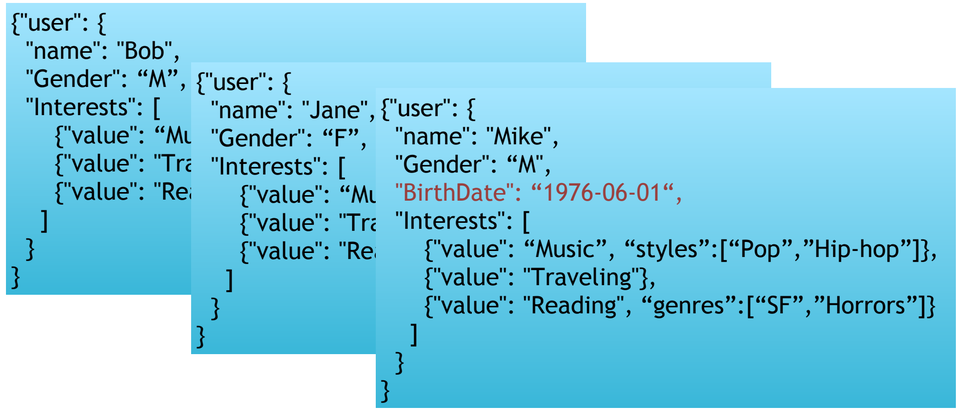
DMA.ClickStream:
-cookie
-URL
-Referer
-Event_type
-Event_dt
-IP
-User_Agent
-.....

100m -> 1b ->10b visits per day

Manychat

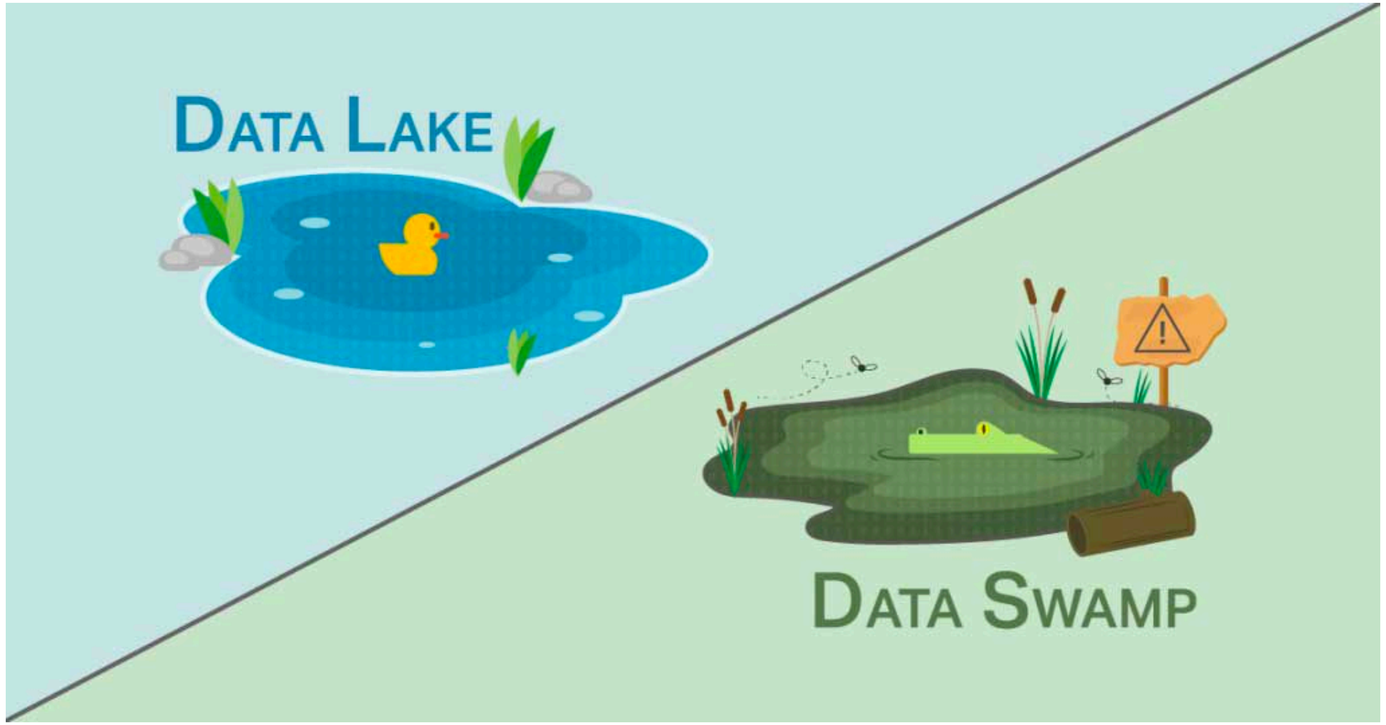
Clickstream в JSON, проблемы:

1. Колодочные СУБД не ускоряют SQL….
2. Схема: какие атрибуты использовать в наших JSON?
3. Безопасны ли JSON в плане перс-данных?



Manychat

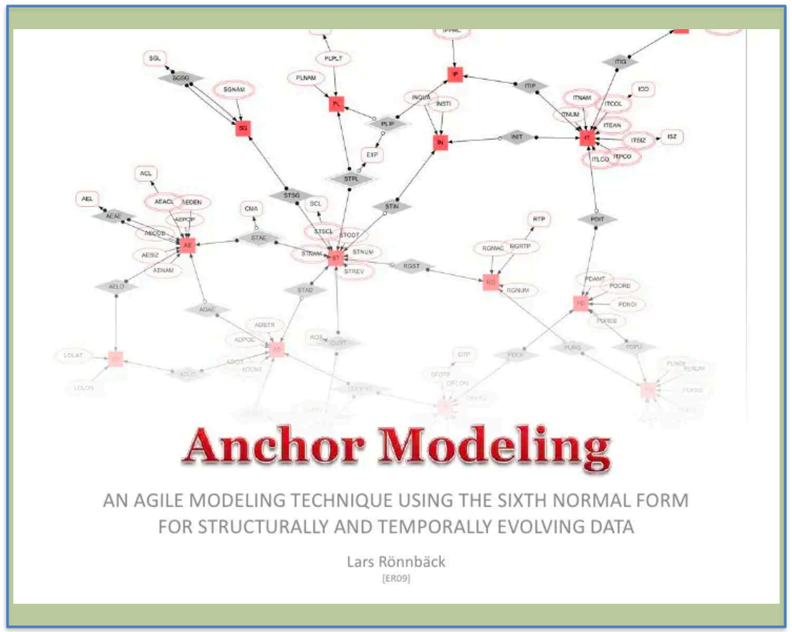
Clickstream в JSON:



Можно, но не долго…

Я фанат Anchor Modeling

Manychat

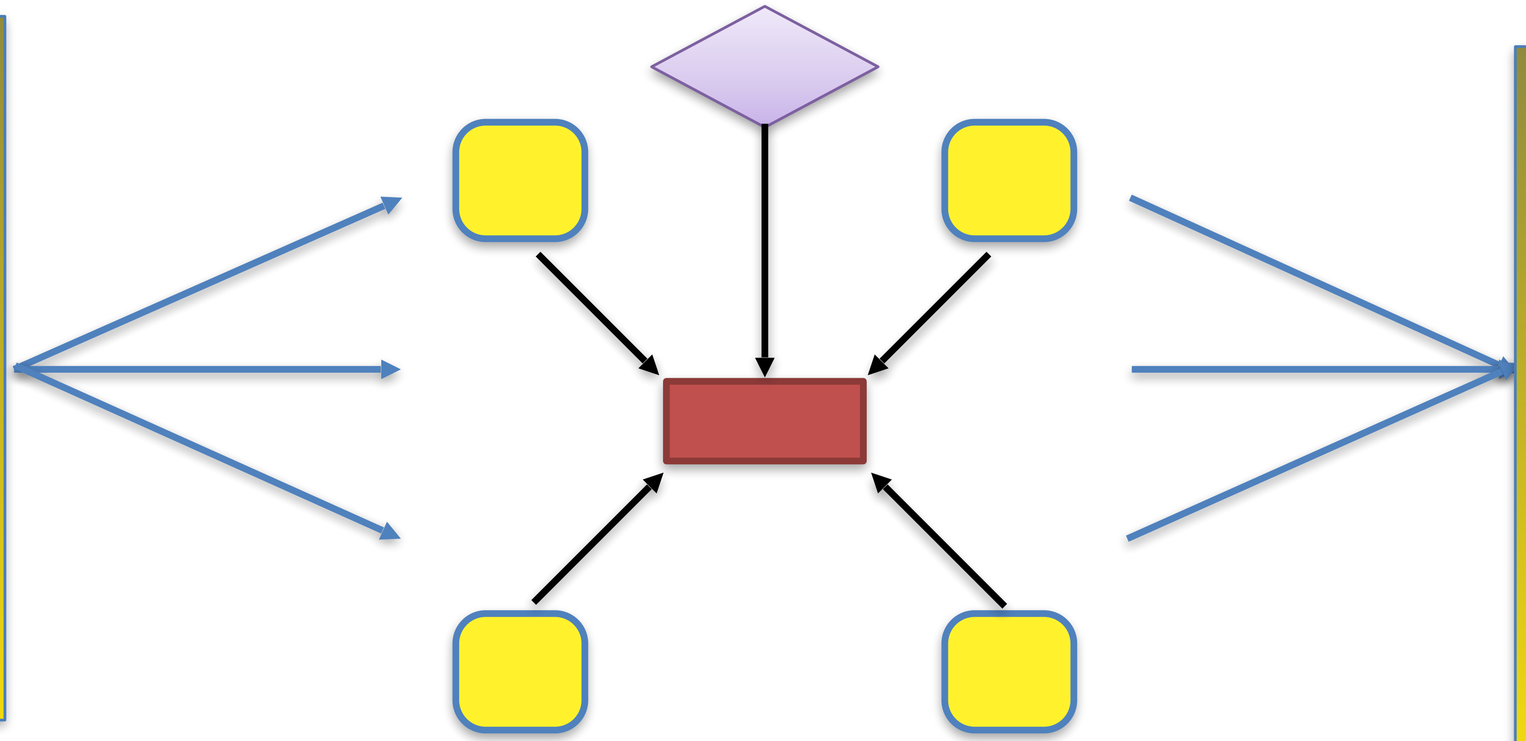


- **Anchor Modeling** - развитие **Data Vault**
- **6NF**
- Таблица для сущности - Anchor
- Таблица для атрибута - Attribute
- Таблица для связи - Tie

Manychat

Нормализованный clickstream

Event:
-cookie
-URL
-Referer
-Event_type
-Event_dt
-IP
-User_Agent
-.....



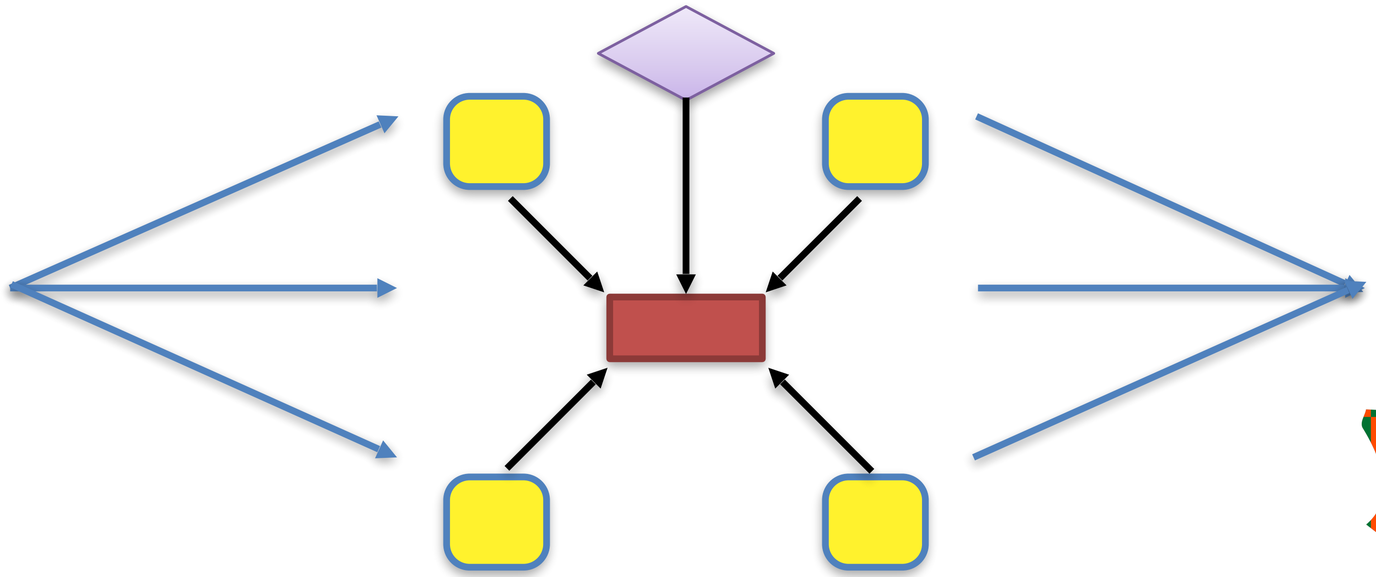
DMA.ClickStream:
-cookie
-URL
-Referer
-Event_type
-Event_dt
-IP
-User_Agent
-.....

100m -> 1b ->10b visits per day

Clickstream в 6NF, проблемы:

Manychat

1. Сотни гигантских таблиц.
2. Для загрузки - JOIN миллиардов строк.
3. Для анализа - JOIN миллиардов строк.



Clickstream в StarSchema?

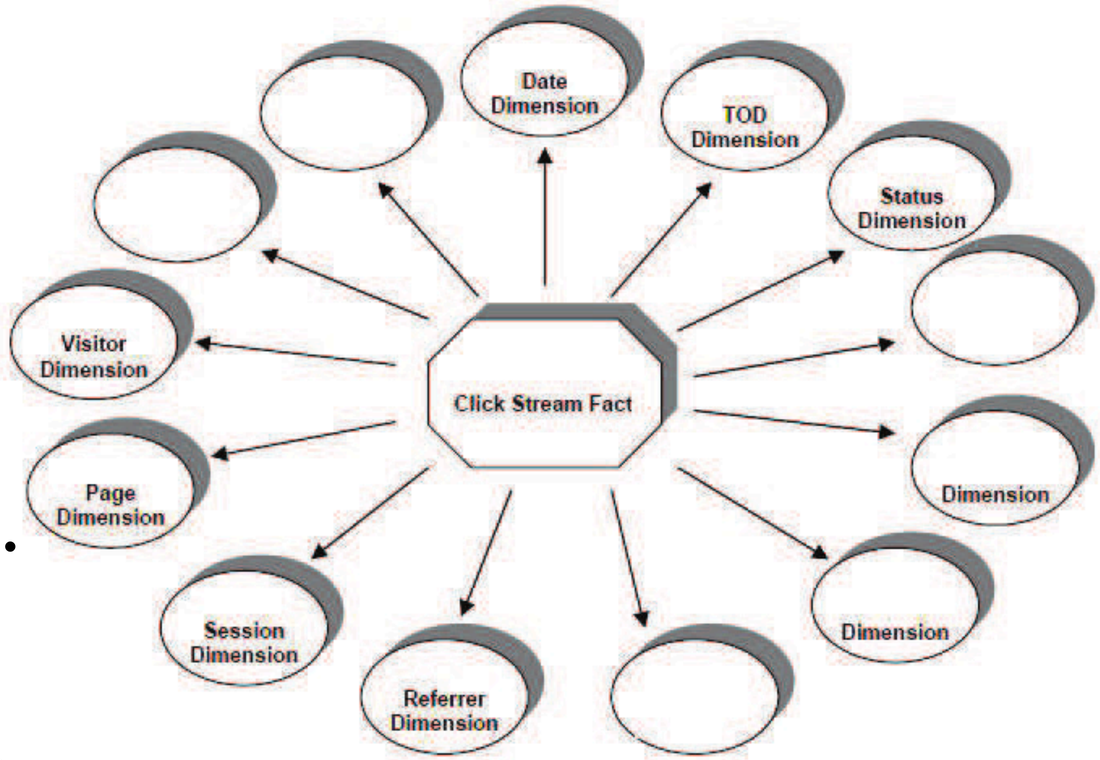
Manychat

1. Фактовая таблица в сотни столбцов. ClickHouse?

2. Разреженная - большая часть таблицы пуста.

3. Для анализа нужны JOIN fact*dimension. ClickHouse?

4. Каждое событие с парой атрибутов -> +2 столбца для всей таблицы.



Решение часть 01: Одна таблица с основными (!) атрибутами ~ Activity Schema

Manychat

Stream_id	User_id	event_id	action_datetime
320003401	42	11	2021-04-23 13:11:48
320003402	51	1	2021-04-27 11:36:03
320003403	42	7	2021-07-21 14:59:01
320003404	23	7	2022-08-05 10:32:58

1. Быстрая и дешевая загрузка.
2. Колоночность. Компрессия. ClickHouse!
3. Удобная базовая аналитика:
    1. DAU/WAU/MAU
    2. Funnels.
    3. Conversions.

Manychat

How many customers per day?

```
SELECT
    TS::date,
    count(DISTINCT CUSTOMER)
FROM S
GROUP BY 1
```

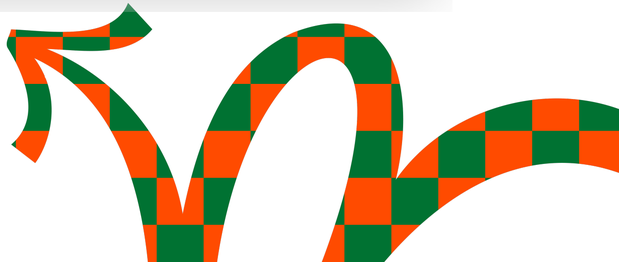
TS	ACTIVITY	CUSTOMER	...
2021-02-02	**Completed Order**	lamond.patin18@example.com	
2021-02-02	**Viewed Page**	lamond.patin18@example.com	
2021-02-02	**Submitted Ticket**	cyndia.atlee@example.com	
2021-02-03	**Viewed Page**	cyndia.atlee@example.com	
2021-02-04	**Viewed Page**	bradford.h@example.com	
2021-02-10	**Opened Email**	bradford.h@example.com	
2021-02-11	**Received Product**	rissa.ailey1908@example.com	

Manychat

Show me the customer funnel?

```
SELECT
    Activity,
    count(DISTINCT CUSTOMER)
FROM S
GROUP BY 1
ORDER BY 2 Desc
```

TS	ACTIVITY	CUSTOMER	...
2021-02-02	**Completed Order**	lamond.patin18@example.com	
2021-02-02	**Viewed Page**	lamond.patin18@example.com	
2021-02-02	**Submitted Ticket**	cyndia.atlee@example.com	
2021-02-03	**Viewed Page**	cyndia.atlee@example.com	
2021-02-04	**Viewed Page**	bradford.h@example.com	
2021-02-10	**Opened Email**	bradford.h@example.com	
2021-02-11	**Received Product**	rissa.ailey1908@example.com	



Manychat

Show me the top buyers?

```
SELECT
     CUSTOMER,
     count(DISTINCT TS)
FROM S
WHERE ACTIVITY='Received product'
ORDER BY 2 DESC
LIMIT 10
```

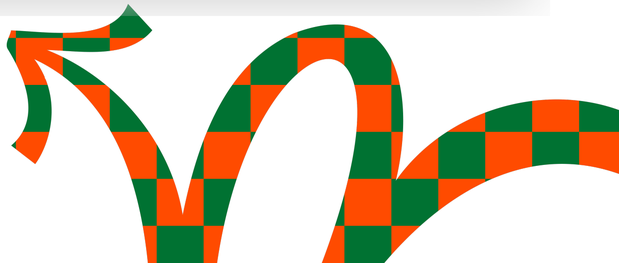
TS	ACTIVITY	CUSTOMER	...
2021-02-02	**Completed Order**	lamond.patin18@example.com	
2021-02-02	**Viewed Page**	lamond.patin18@example.com	
2021-02-02	**Submitted Ticket**	cyndia.atlee@example.com	
2021-02-03	**Viewed Page**	cyndia.atlee@example.com	
2021-02-04	**Viewed Page**	bradford.h@example.com	
2021-02-10	**Opened Email**	bradford.h@example.com	
2021-02-11	**Received Product**	rissa.ailey1908@example.com	

Manychat

Show me customers, who forgot to order?

```
SELECT * FROM (
      SELECT
       CUSTOMER,
       LAST(ACTIVITY)
        OVER(PARTITION BY CUSTOMER
          ORDER BY TS) AS LA
      FROM S
) T WHERE LA = 'Product selected'
```

TS	ACTIVITY	CUSTOMER	...
2021-02-02	**Completed Order**	lamond.patin18@example.com	
2021-02-02	**Viewed Page**	lamond.patin18@example.com	
2021-02-02	**Submitted Ticket**	cyndia.atlee@example.com	
2021-02-03	**Viewed Page**	cyndia.atlee@example.com	
2021-02-04	**Viewed Page**	bradford.h@example.com	
2021-02-10	**Opened Email**	bradford.h@example.com	
2021-02-11	**Received Product**	rissa.ailey1908@example.com	



Решение часть 02: нормализованные таблицы с деталями

Manychat

Stream_id	device_type
320003403	desktop
320003404	tablet

Stream_id	IP
320003401	126.035.201.003

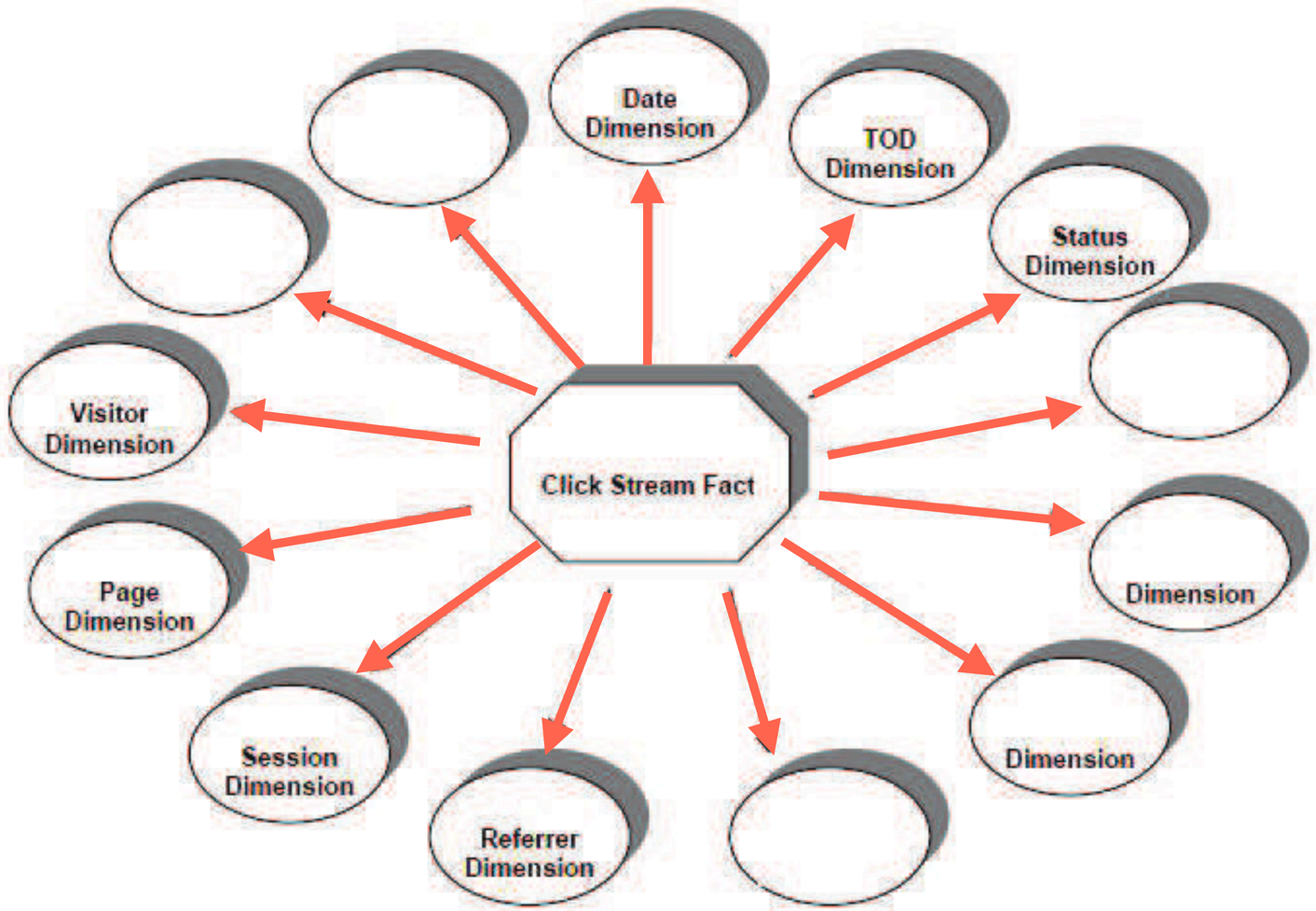
Stream_id	Error
320003402	404

Stream_id	User_id	event_id	action_datetime
320003401	42	11	2021-04-23 13:11:48
320003402	51	1	2021-04-27 11:36:03
320003403	42	7	2021-07-21 14:59:01
320003404	23	7	2022-08-05 10:32:58

1. Таблицы деталей маленькие
2. Проще наполнять/использовать
3. Если не нужны - игнорируй

Это разве не StarSchema?

Manychat



Решение часть 02: нормализованные таблицы с деталями

Manychat

Stream_id	device_type
320003403	desktop
320003404	tablet

Stream_id	IP
320003401	126.035.201.003

Stream_id	Error
320003402	404

Stream_id	User_id	event_id	action_datetime
320003401	42	11	2021-04-23 13:11:48
320003402	51	1	2021-04-27 11:36:03
320003403	42	7	2021-07-21 14:59:01
320003404	23	7	2022-08-05 10:32:58

1. Таблицы деталей маленькие
2. Проще наполнять/использовать
3. Если не нужны - игнорируй

Решение часть 03:
JSONs в скрытой таблице деталей

Manychat

Stream_id	User_id	event_id	action_datetime
320003401	42	11	2021-04-23 13:11:48
320003402	51	1	2021-04-27 11:36:03
320003403	42	7	2021-07-21 14:59:01
320003404	23	7	2022-08-05 10:32:58

Stream_id	Data
320003402	{name:”someName”, email: “sample@gmail.com, token:”34da32c67a54af”, data:[……..]}

1. JSONs спрятаны
2. Нет регулярного обращения
3. Нет риска утечки персданных
4. JSON и переданные доступны для извлечения.

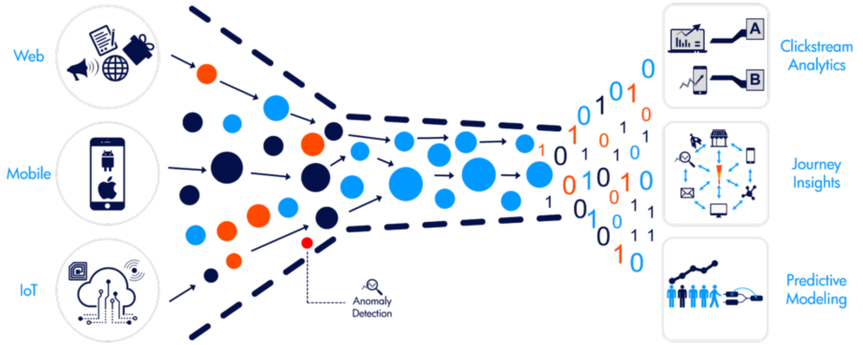
Classification of events:

Manychat

1. We have hundreds/thousands events from many sources.

2. Which events are visit/click/action?

3. Every day we got dozen new events, how to classify them?



Решение часть 04: классификация событий с помощью тегов

Manychat

event_id	event_name	tags
1	fb:message_out_echo:cards	echo,outgoing_from_account,facebook
7	fb:fb_button_click	facebook,contact_action
11	fb:request_error	

1. 25 тегов
2. 4000+ событий
3. Дневной мониторинг

Stream_id	User_id	event_id	action_datetime
320003401	42	11	2021-04-23 13:11:48
320003402	51	1	2021-04-27 11:36:03
320003403	42	7	2021-07-21 14:59:01
320003404	23	7	2022-08-05 10:32:58

Дневной мониторинг событий

Manychat

MCEVENTTYPE_ID	EVENT NAME	CONTACTS	BOTS	PROFILES	VISITS
67651	app:ai:apply	0	3	0	0
67645	app:unpublished_flow_modal_change_checkbox	0	10508	0	0
67644	app:unpublished_flow_modal_left	0	3756	82	0
67646	app:unpublished_flow_modal_return	0	2895	0	0
67647	app:unpublished_flow_modal_shown	0	9779	0	0
67643	contact:deleted_bu_user	6984	0	0	0

Использование тегов в использовании событий

Manychat

```
SELECT
 EVENT_DATE::date,
 count(DISTINCT MCBOTACCOUNT_ID)
FROM S_ContactEvent
JOIN MCEventType_Manifest
USING(MCEventType_id)
WHERE ARRAY_CONTAINS('ai'::VARIANT, tags_raw)
GROUP BY 1
ORDER BY 1 DESC ;
```

1. Filter out just events with needed tags…
2. Join with a stream
3. …

Решение часть 05: Metadata Register - обязательные атрибуты событий

Manychat

event	attribute
visit	event_type
visit	event_date
visit	Cookie
visit	IP
registration	event_type
registration	event_date
registration	Cookie
registration	IP
registration	User_ID

Stream_id	User_id
…	…

Stream_id	type	date	cookie
…	…	…	…
…	…	…	…
…	…	…	…
…	…	…	…

Stream_id	IP
…	…

Решение часть 05: Metadata Register -> UNION схожих событий

Manychat

event	attribute
visit	event_type
visit	event_date
visit	Cookie
visit	IP
registration	event_type
registration	event_date
registration	Cookie
registration	IP
registration	User_ID

```
SELECT
   type, date, cookie, IP
FROM visit
UNION ALL
SELECT
   type, date, cookie, IP
FROM registration
```

Stream_id	type	date	cookie
…	…	…	…
…	…	…	…
…	…	…	…

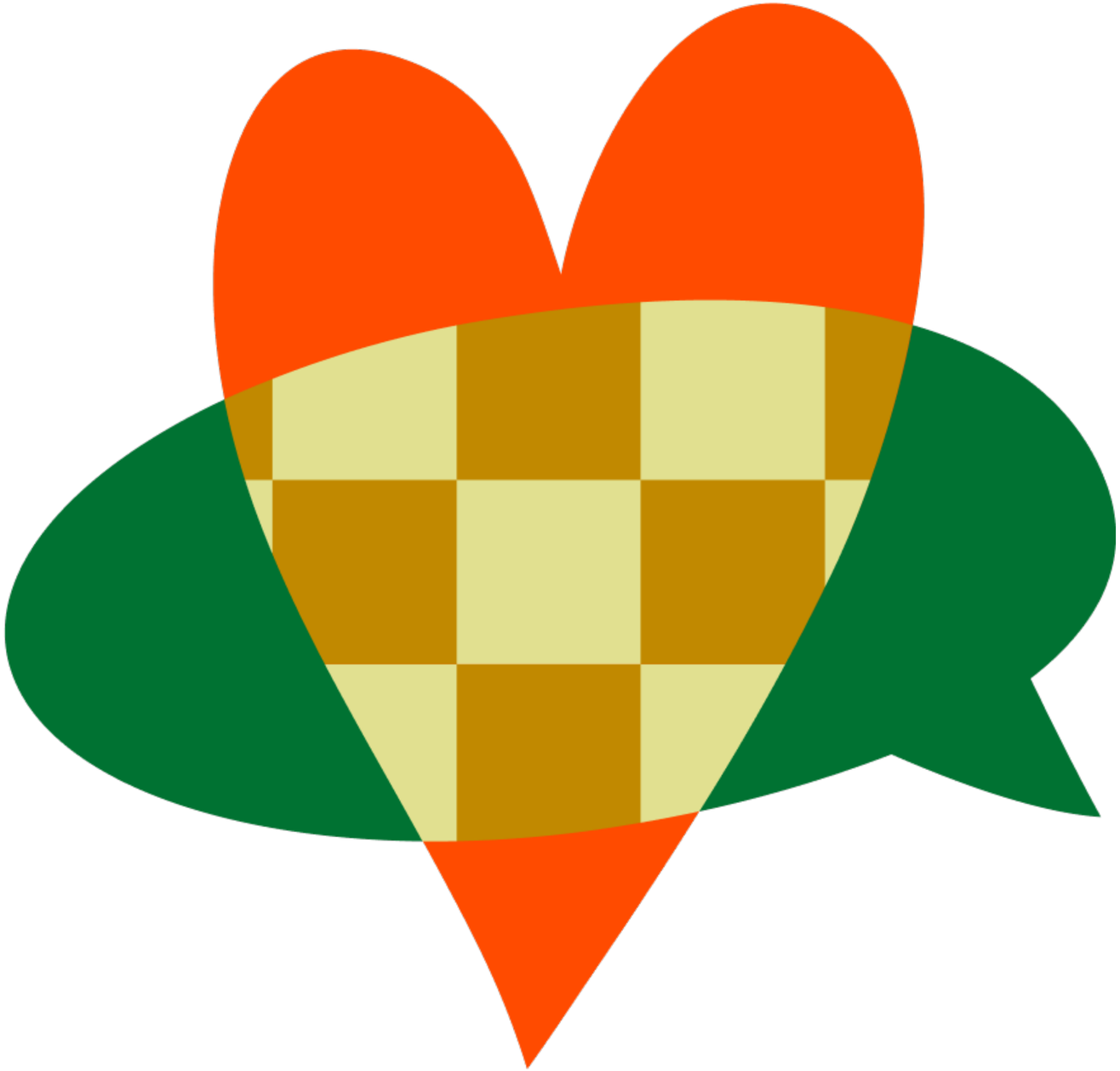
Stream_id	type	date	cookie	IP
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

Статистика

Manychat

TABLE_NAME	ROW_COUNT	DAILY_ROWS	BYTES	SOURCE_SIZE
S_CONTACTEVENT	84,968,403,568	65 millions	1,293,047,520,256	6Tb
S_CONTENTPROCESSING	35,277,674,882	35 millions	548,707,631,616	2.5Tb
S_BOTACCOUNTEVENT	12,701,601,251	15 millions	147,139,272,704	0.7Tb
S_MCPIXELEVENT	8,860,669,610	8 millions	123,978,569,216	0.5Tb
S_PROFILEEVENT	229,635,007	0.5 millions	2,746,594,304	12Gb
S_VISITOREVENT	90,042,705	90.000	1,120,123,904	5Gb
S_IGACCOUNTEVENT	20,092,488	14.000	252,306,944	1Gb

Thank You



Nikolai Golov
nikolay@manychat.com