avito.tech

Москва — 2023

ЧТО ДЕЛАТЬ, ЕСЛИ

ХРАНИЛИЩЕ РАСТЁТ СЛИШКОМ БЫСТРО

Александр Филатов

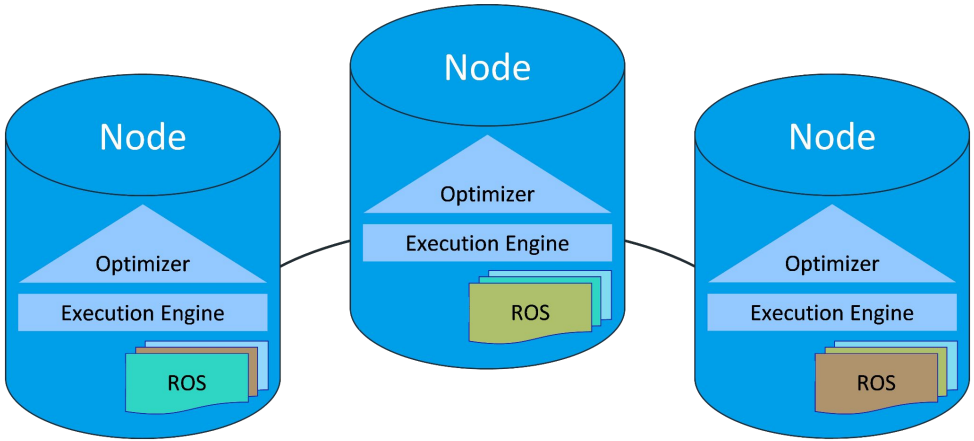
Лид команды Integration

Интро

КТО МЫ ТАКИЕ И ГДЕ ХРАНИМ ДАННЫЕ

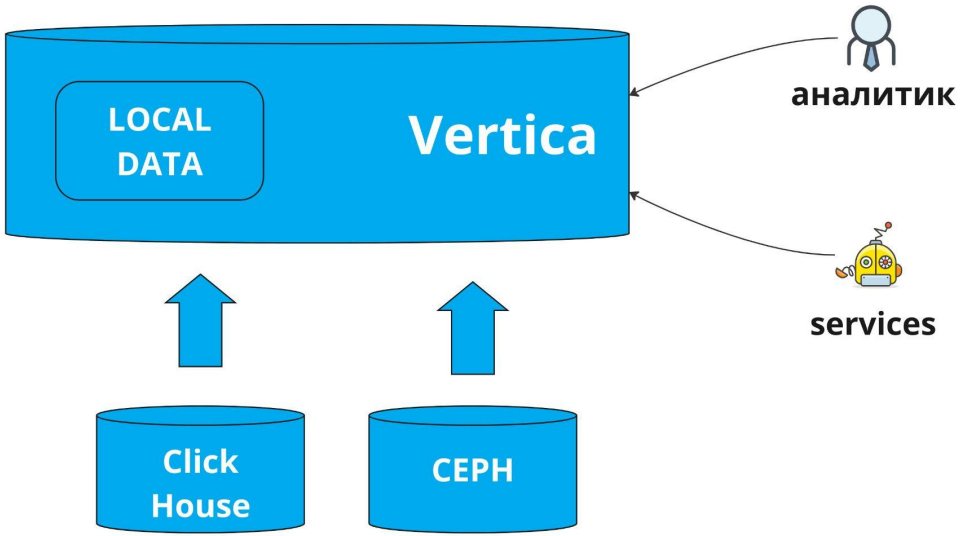
Vertica

1PB data
50 nodes

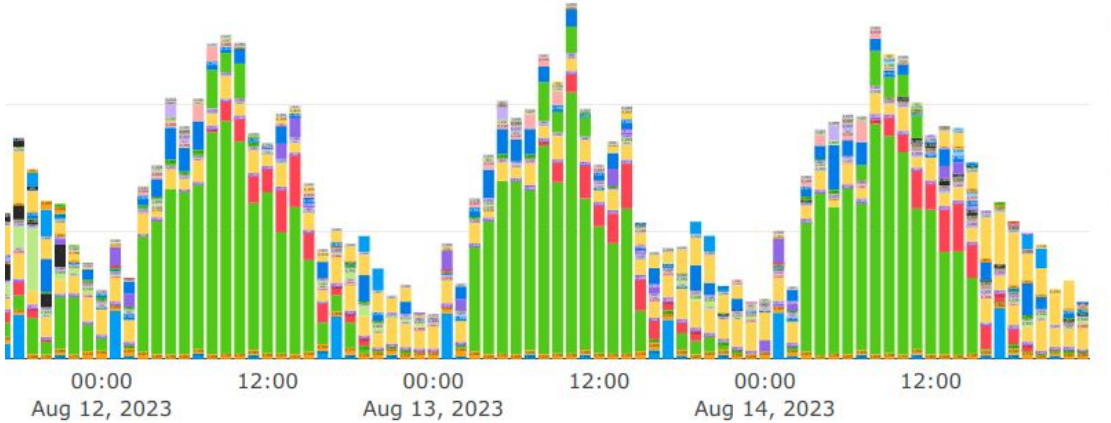


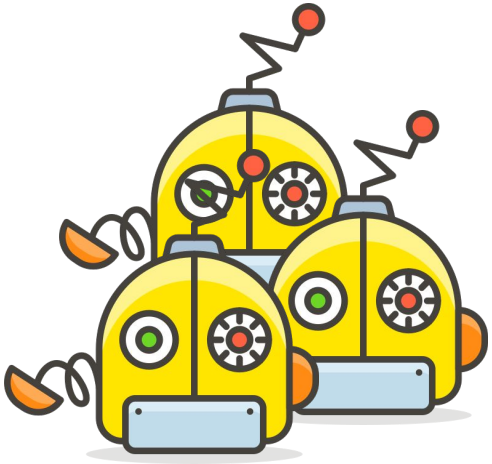
Vertica+

150 int. srv
200 dau
1M queries

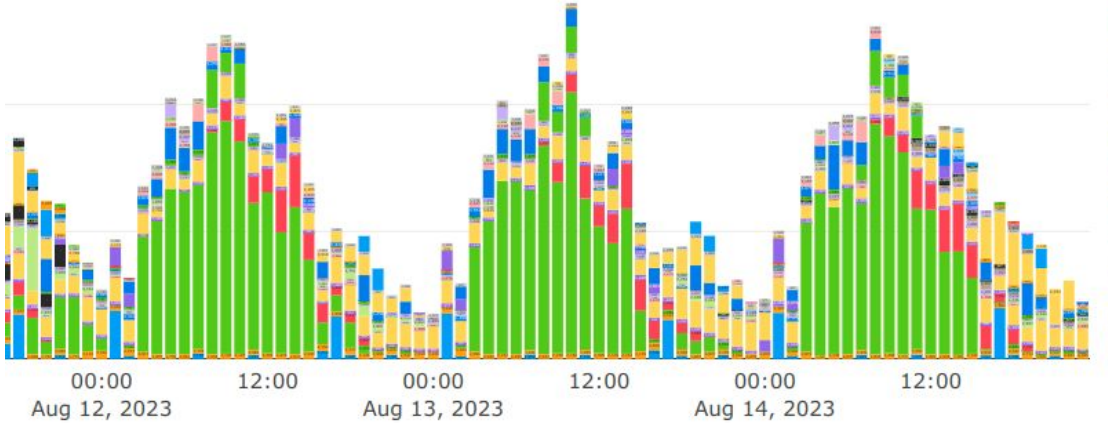


cpu cycles per hour

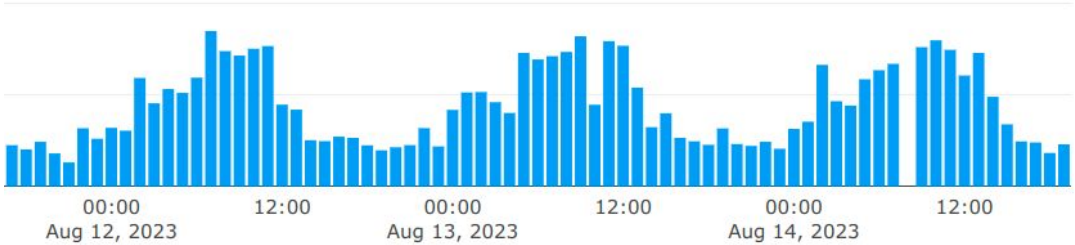


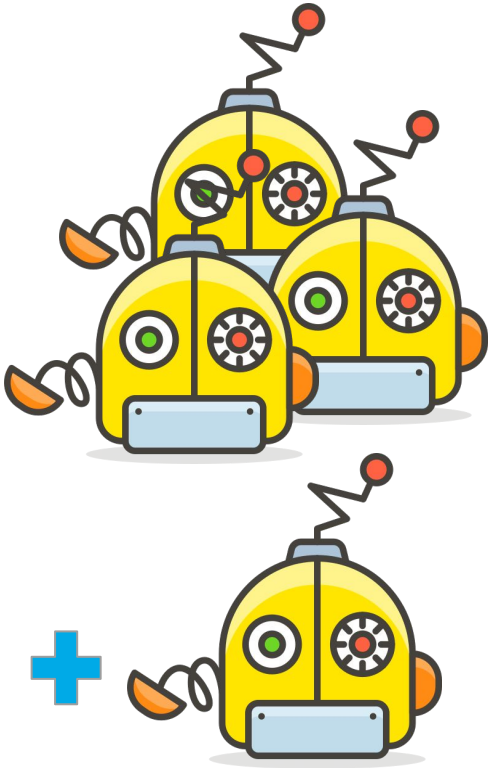


cpu cycles per hour



query duration



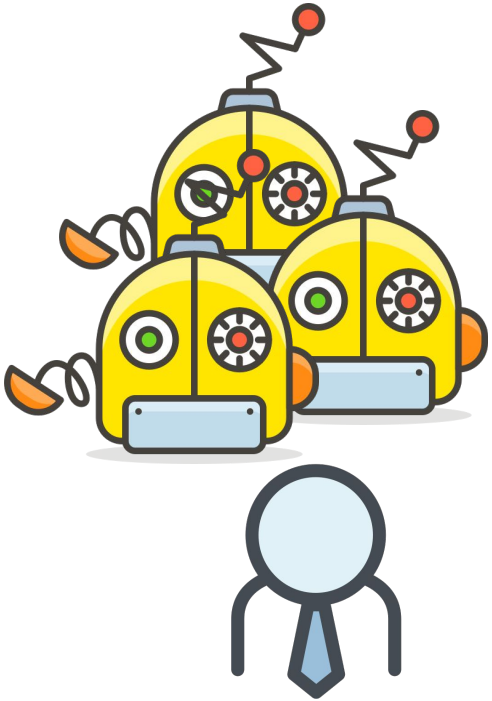


Проблема 1:

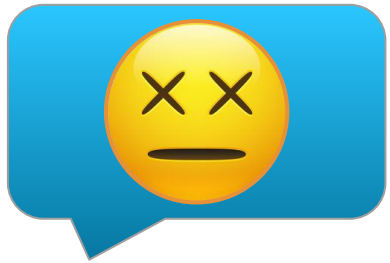
«Борьба за ресурсы»

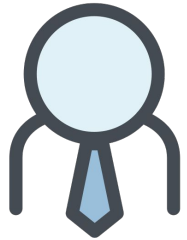
Ожидание готовности:

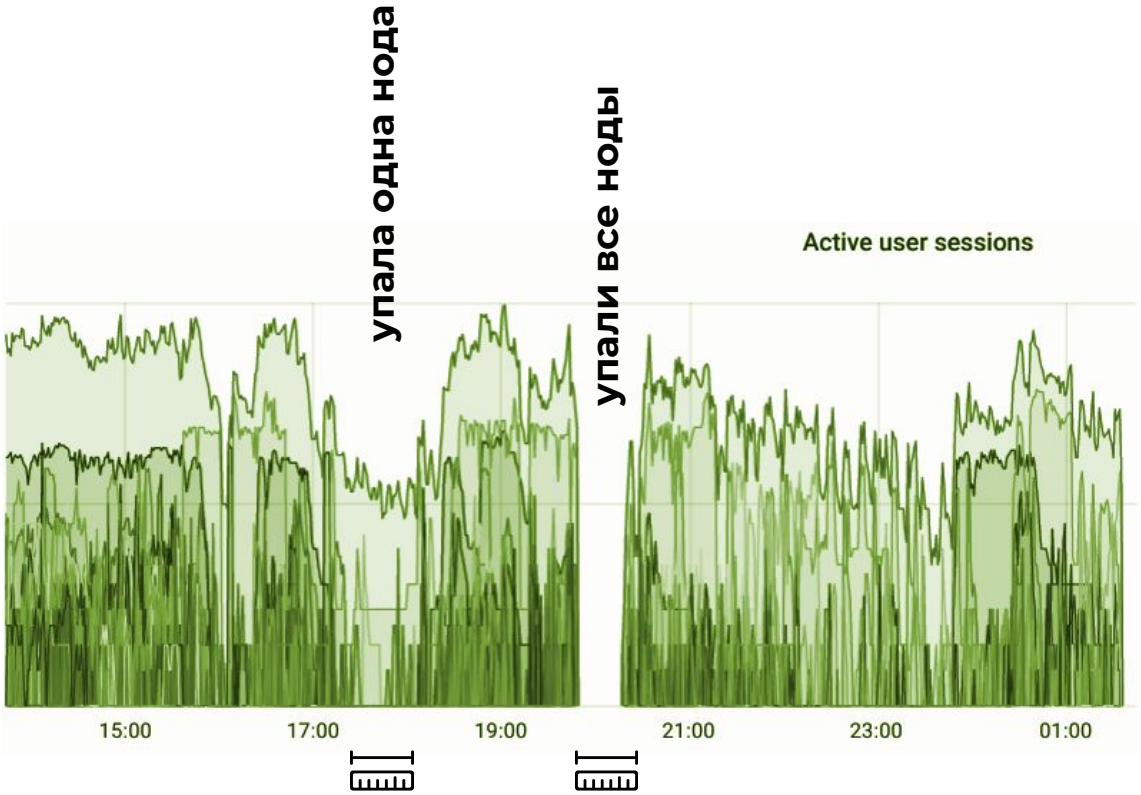
- 8AM
- 10AM
- 12AM
- during the day
- one day delay is ok



```
create table tmp_arrays
(
    a array[array[int]]
);
```





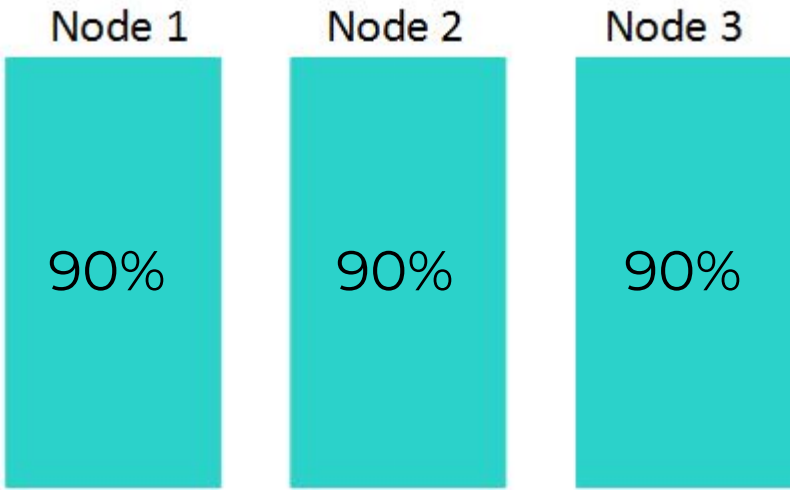


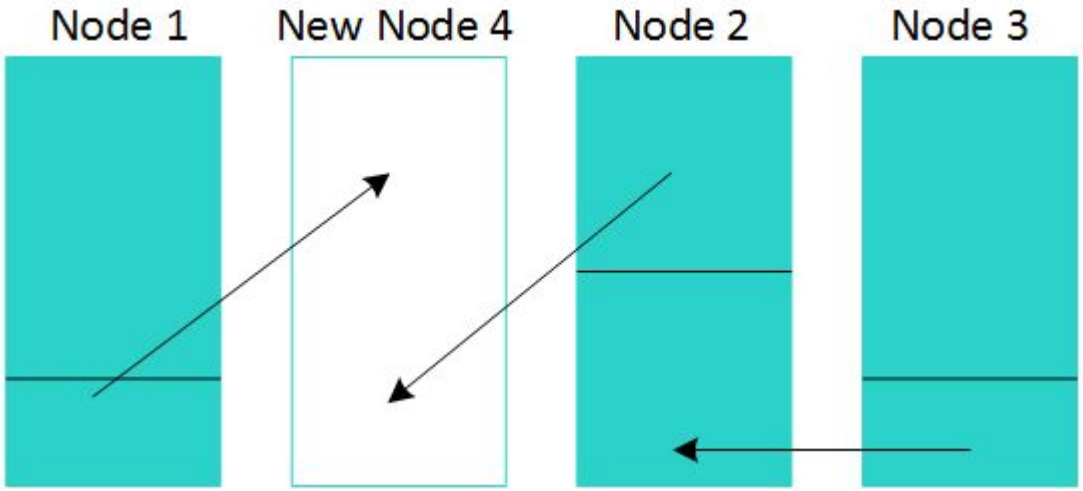
восстановление = **45мин !!!**

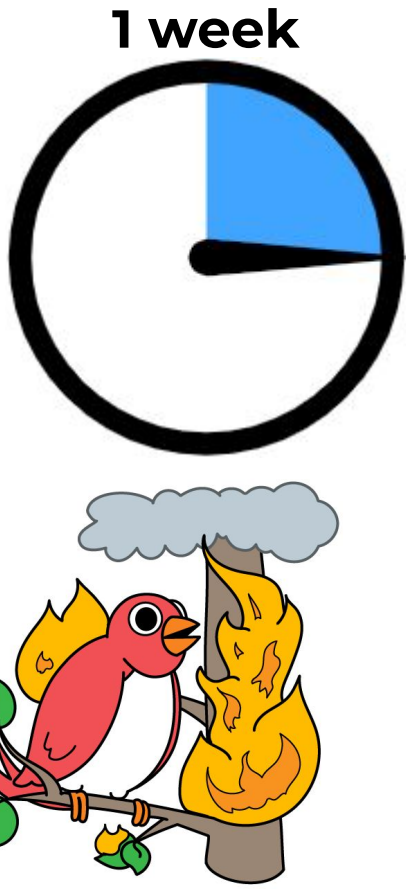
```
Connection closed

Node failure during
execution

Insufficient
projections to
answer query

Cannot modify
temporary table
because a node has
recovered
```

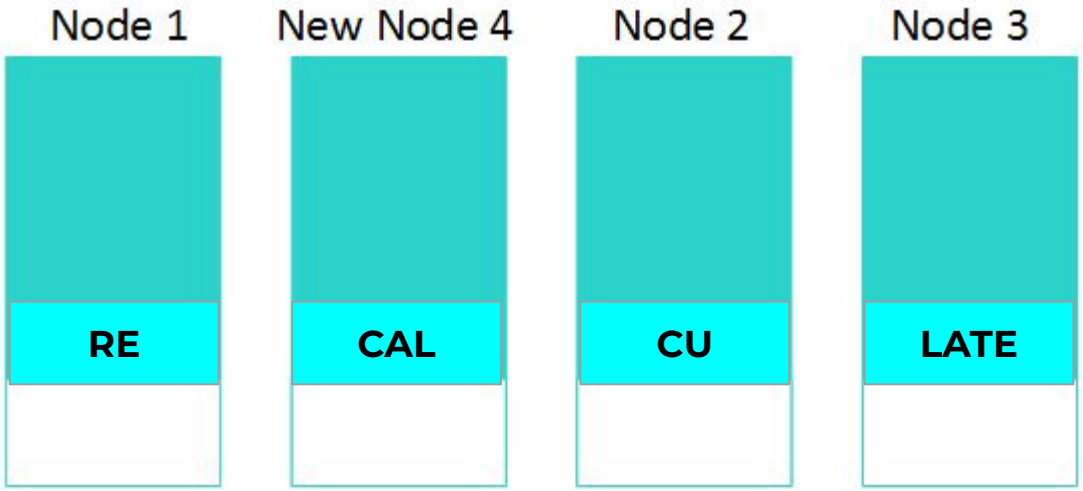
Проблема 2::

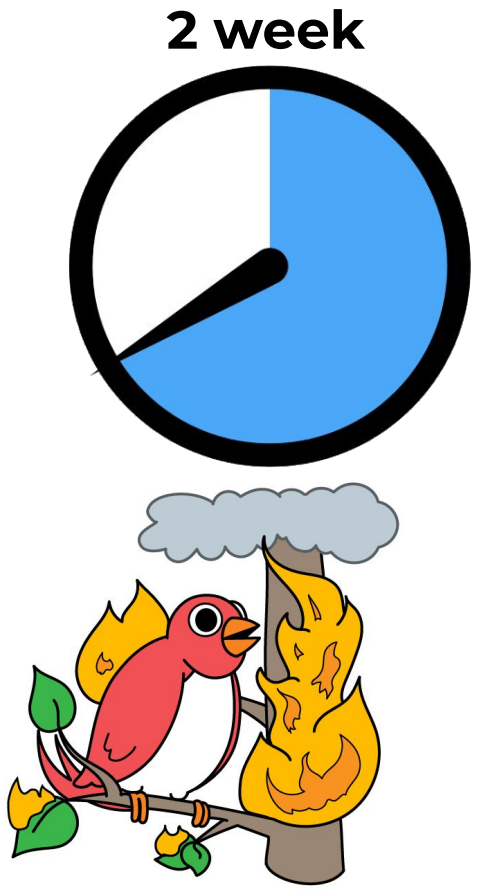
«Эффект домино»


Node 1
Node 2
Node 3
90%
90%
90%


Node 1
New Node 4
Node 2
Node 3

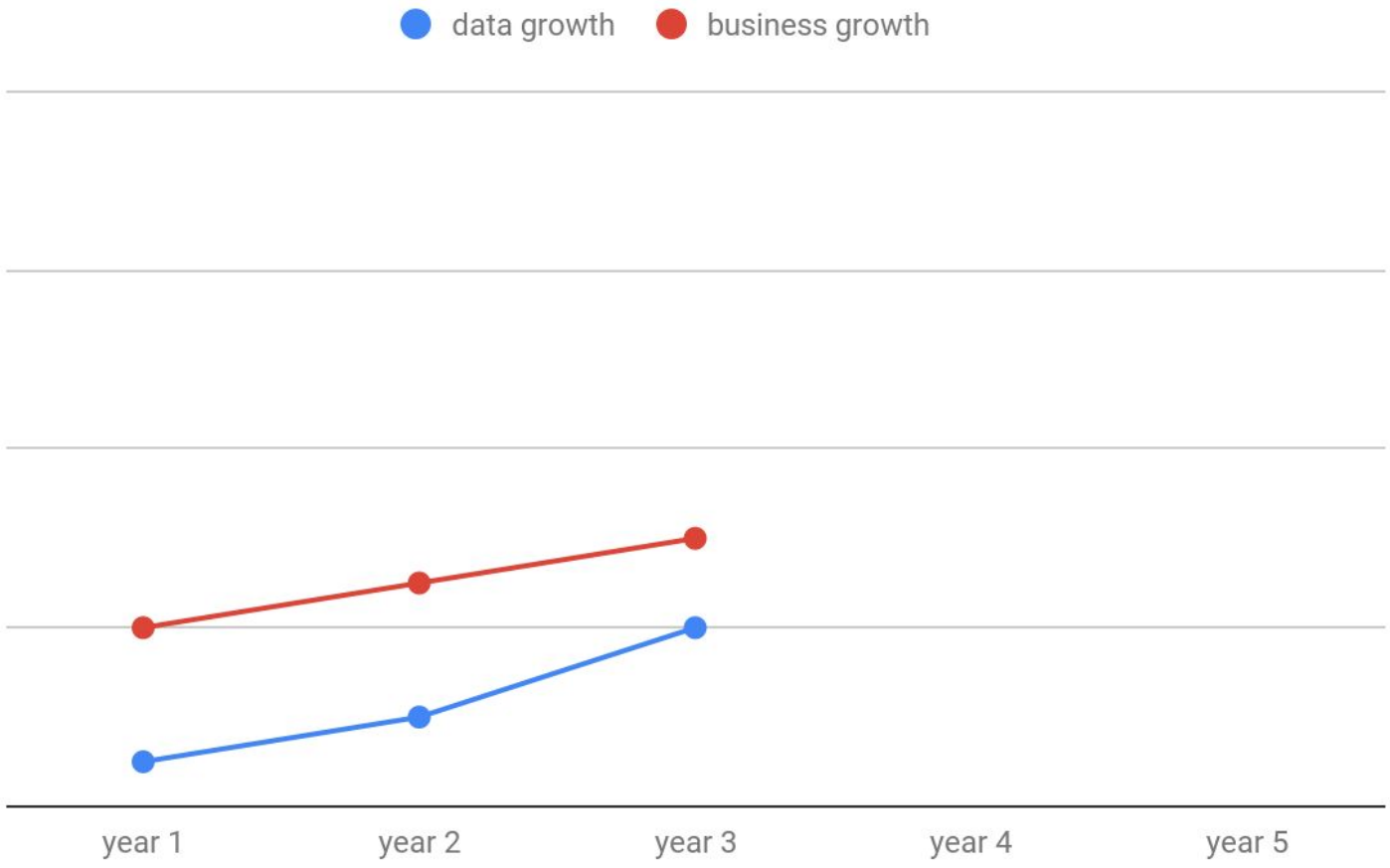

1 week

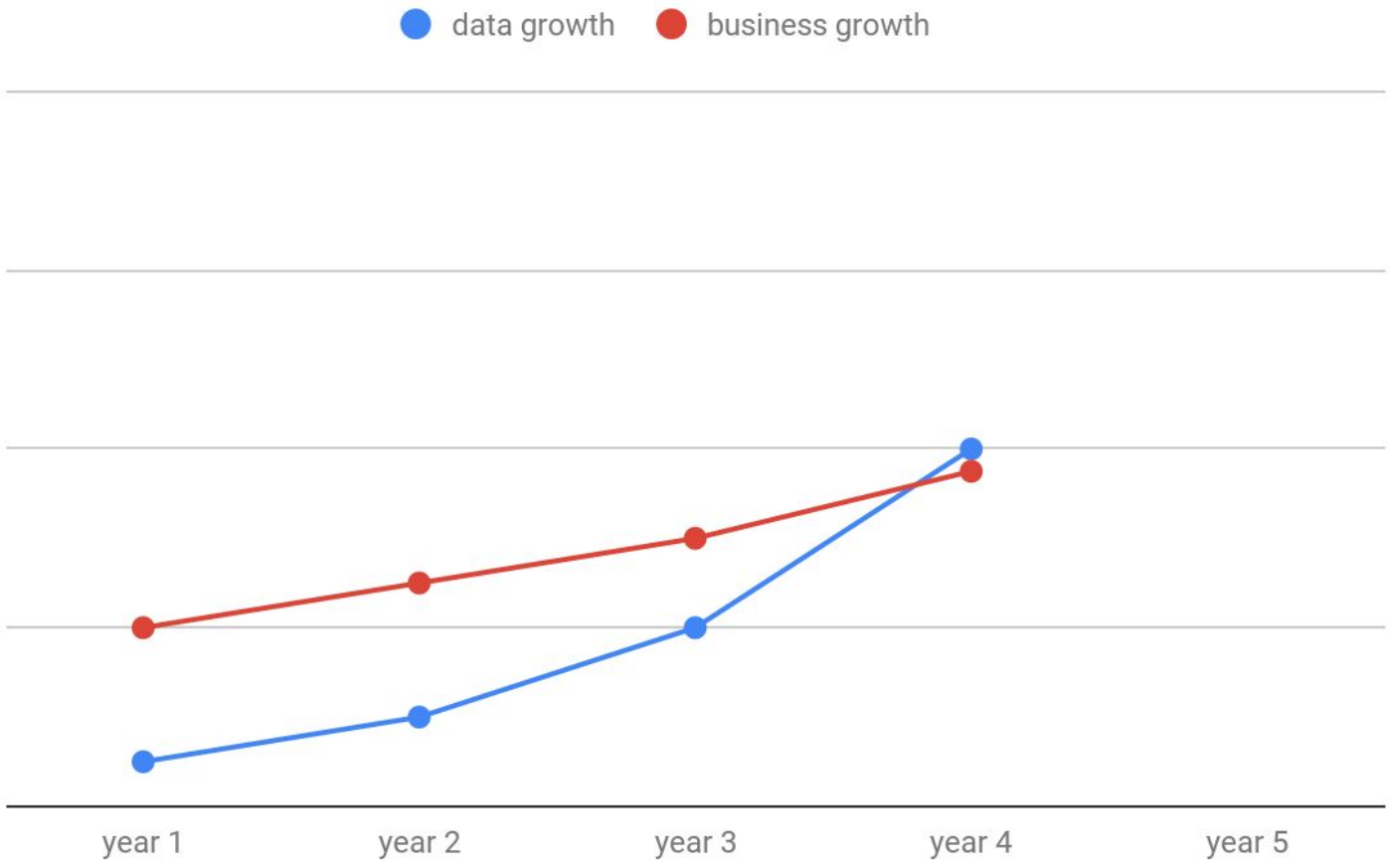

Node 1
New Node 4
Node 2
Node 3
RE
CAL
CU
LATE

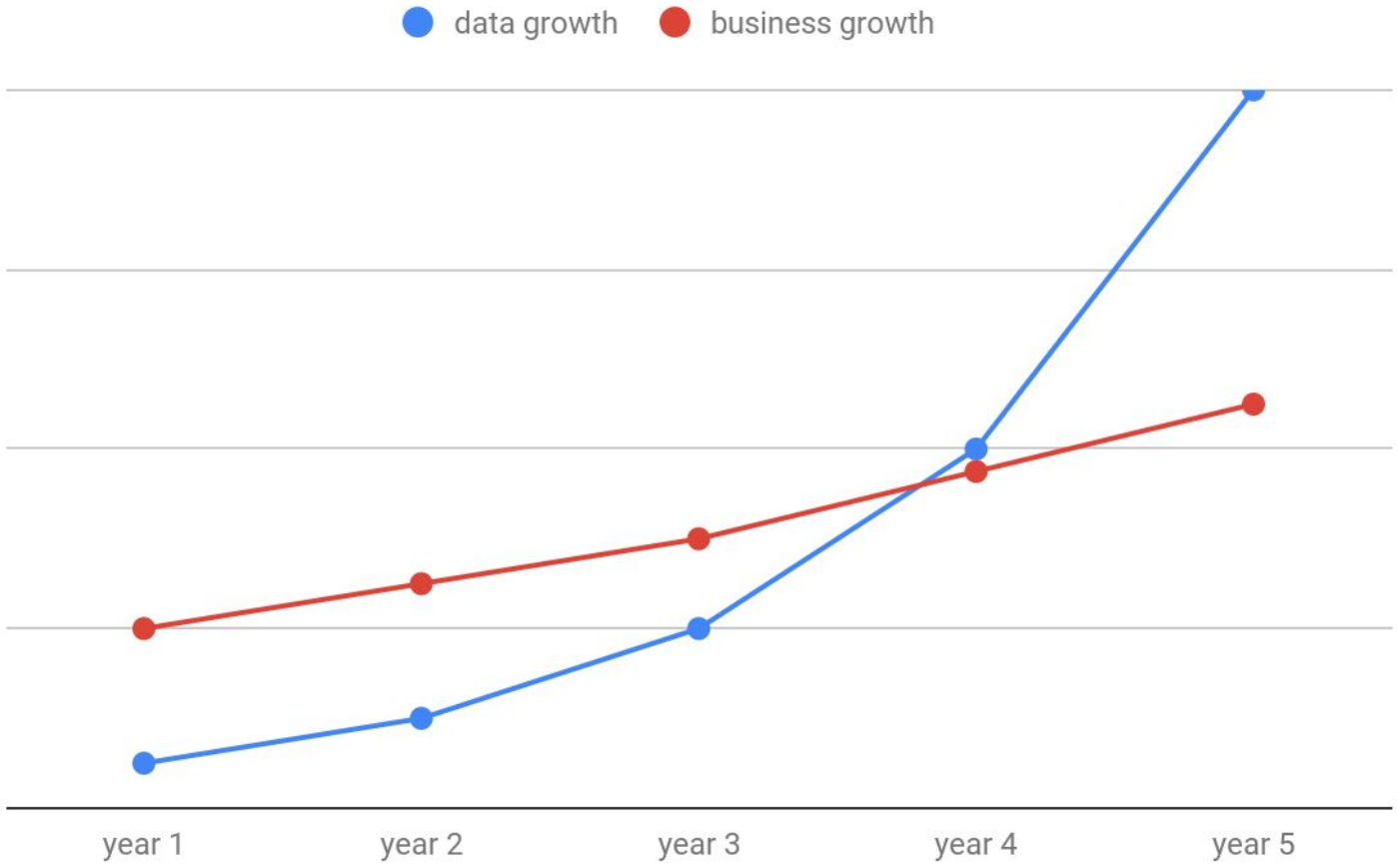

2 week

Проблема 3:

«Немасштабируемость»


data growth
business growth
year 1
year 2
year 3
year 4
year 5


data growth
business growth
year 1
year 2
year 3
year 4
year 5


data growth
business growth
year 1
year 2
year 3
year 4
year 5

Проблема 4:

«Стоимость»

ЧТО МЫ ИМЕЕМ НА НАЧАЛО 2022

Проблемы:

- Борьба за ресурсы мешает давать гарантии
- Эффект домино мешает продуктовым интеграциям
- Масштабируемость по месту, а не потребности в расчётах

Скорость роста данных

Усугубление

ЧТО МЫ ИМЕЕМ НА НАЧАЛО 2022

Проблемы:

- Борьба за ресурсы мешает давать гарантии
- Эффект домино мешает продуктовым интеграциям
- Масштабируемость по месту, а не потребности в расчётах

Скорость роста данных

Усугубление

Что мы хотим:

- Повысить стабильность работы платформы
- Масштабироваться от вычислений, а не от данных

О ЧЁМ НЕ БУДЕМ ГОВОРИТЬ

- Сокращать нагрузку.
- Поднимать реплики.
- Менять модель данных.
- Распиливать данные.

ПЛАН: БЫСТРЕЕ, ДЕШЕВЛЕ, МАСШТАБИРУЕМЕЕ

Что мы смотрели:

- какие референсы;
- как выбирали;
- почему не пугает.

С чем столкнулись:

- как выгружать;
- как переписать;
- как оптимизировать;
- как не потерять.

Так ли хорошо получилось:

- где мы сейчас;
- устраивает ли скорость;
- достигли ли целей.

1

2
С чем столкнулись

3
Так ли хорошо получилось

ЧТО МЫ СМОТРЕЛИ

- Какие референсы
- Как выбирали
- Почему не пугает

Vertica EON

GreenPlum

Spark

Gluten DataBricks

GreenPlum

Spark

Gluten DataBricks

Presto Velox

Trino

GreenPlum


Presto
Velox
Spark
Trino
Gluten
DataBricks
GreenPlum
China Gartner
GausDB
TiDB
AliORC

StarRocks

Dremio

Presto

Velox

Spark

Trino

Gluten

DataBricks

GreenPlum

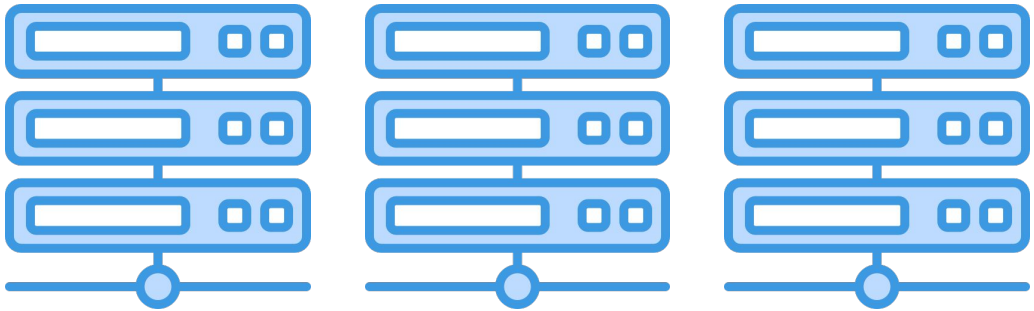
China Gartner

GausDB

TiDB

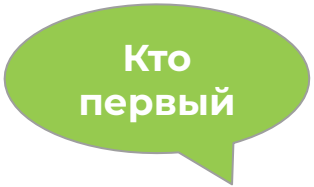
AliORC

НА ЧЁМ ТЕСТИРОВАЛИ

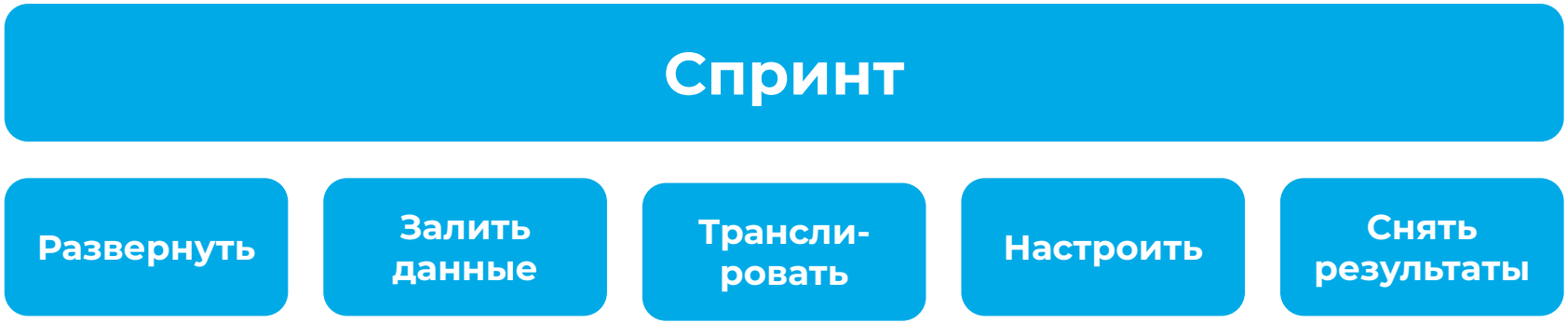


20x

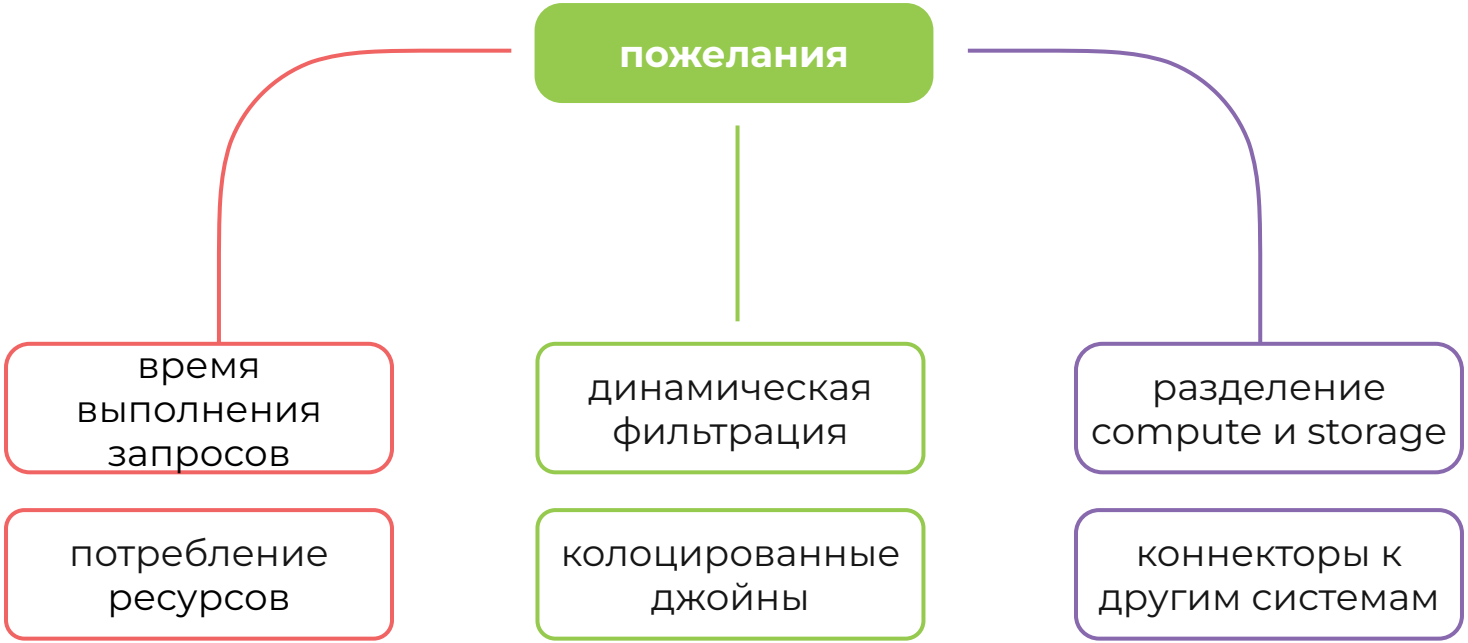
Dell Inc. PowerEdge R740×d 2U
2x Intel(R) Xeon(R) Gold 6240
CPU @ 2.60GHz
72CPU 512RAM 10GbE



КАК ТЕСТИРОВАЛИ



НА ЧТО ОБРАЩАЛИ ВНИМАНИЕ

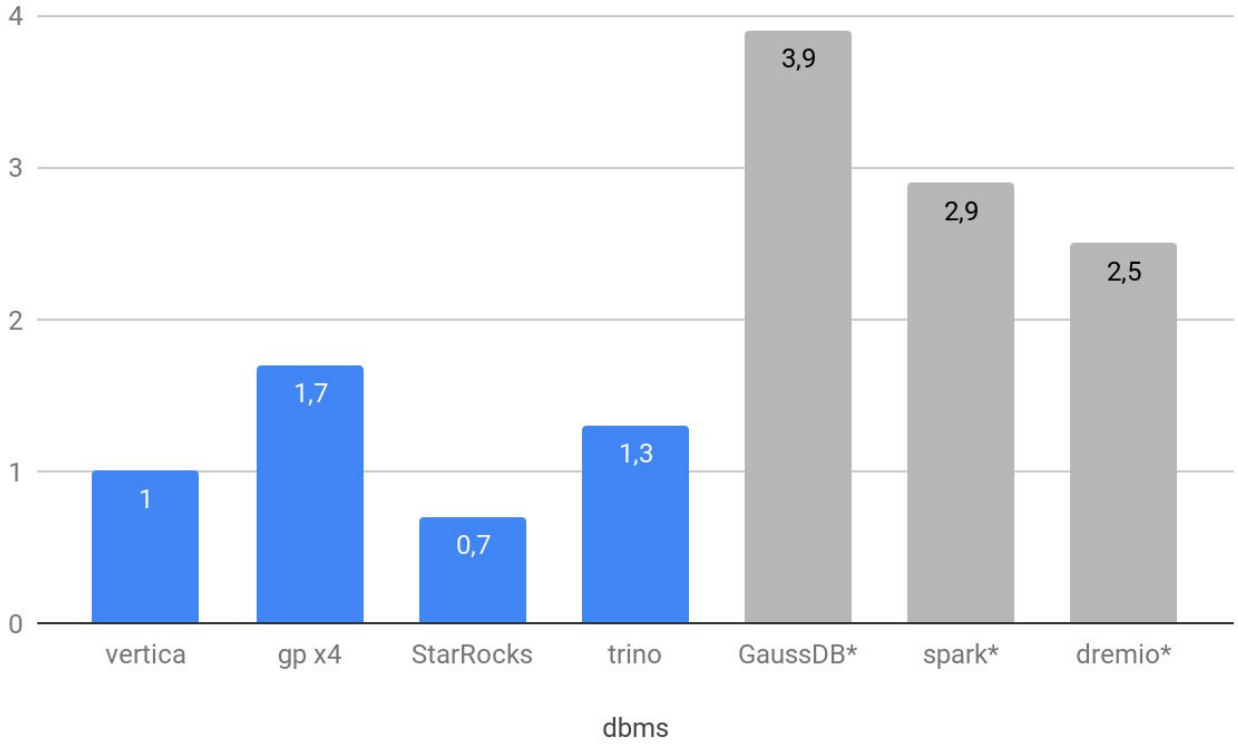


ЧТО СЧИТАЛИ

- На входе узкие таблицы.
- Число строк 100M-1B.
- Множество джойнов.

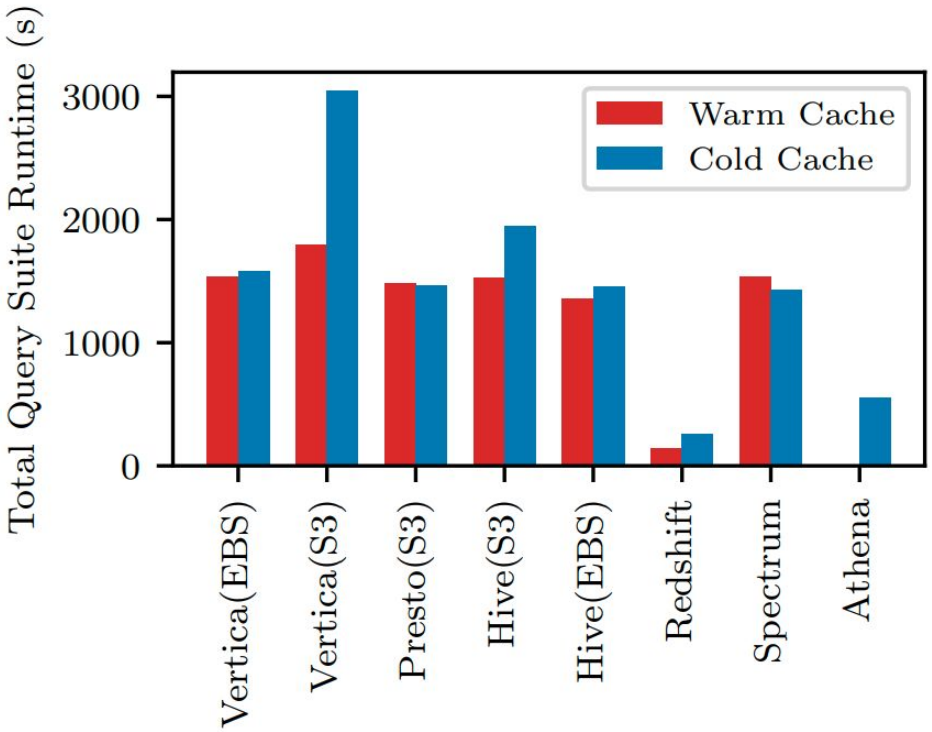
```
insert into dev_DMA.moder_reason
SELECT
    ae.AdmLog_id,
    ra.ModerReason_id,
    ma.ModerAction_id,
    fcn.FraudCode_id,
    concat(fcn.Name, CASE
        WHEN fcn.isPlaceholder THEN (', %s='::varchar(5) || coalesce(regxp.ParameterName, ''::varchar))
        ELSE ''::varchar
    END) AS FraudReason,
    :launch_id
FROM ((((((DDS.L_AdmLog_AdmEvent ae
JOIN DDS.S_AdmLog_EventDate USING(admlog_id))
LEFT JOIN DDS.L_ModerAction_AdmLog ma USING(admlog_id))
LEFT JOIN DDS.L_ModerReason_ModerAction ra USING(moderaction_id))
LEFT JOIN DDS.L_ModerReason_FraudCode mrfc USING(moderreason_id))
LEFT JOIN (
    SELECT t.FraudCode_id,
        t.Name,
        t.isPlaceholder
    FROM (
        SELECT scn.FraudCode_id,
            scn.Name,
            row_number() OVER(PARTITION BY scn.FraudCode_id ORDER BY scn.Actual_date DESC) AS rn,
            ((hc.External_ID = 175) AND (instr(scn.Name, '%s'::varchar(2), 1, 1) > 0))
             AS isPlaceholder
        FROM (DDS.S_FraudCode_Name scn
        JOIN DDS.H_FraudCode hc USING(FraudCode_id))
    ) t
    WHERE (t.rn = 1)
) fcn ON ((fcn.FraudCode_id = mrfc.FraudCode_id)))
LEFT JOIN (
    SELECT ModerAction_id, Name AS ParameterName
    FROM (( (
        SELECT S_ModerAction_ReasonExt.ModerAction_id,
            (max(substr(regexp_substr(replace(S_ModerAction_ReasonExt.ReasonExt, '"'::varchar(1), ''::varchar),
E'175\\:\\[\\d++'::varchar(11), 1, 1, ''::varchar, 0), 6)))::int AS external_id
        FROM DDS.S_ModerAction_ReasonExt
        GROUP BY S_ModerAction_ReasonExt.ModerAction_id
    ) t
    JOIN DDS.H_CategoryParameter USING(External_id))
    JOIN (
        SELECT *, ROW_NUMBER() OVER(PARTITION BY CategoryParameter_id ORDER BY Actual_date DESC) as rn
        FROM DDS.S_CategoryParameter_Name
    ) n USING(CategoryParameter_id))
    WHERE rn = 1
) regxp USING(ModerAction_id))
WHERE ae.AdmEvent_id = ANY (ARRAY[7, 50, 70])
  and (EventDate >= now() - 730 AND ae.AdmEvent_id = ANY (ARRAY[7, 50, 70]) OR ae.AdmEvent_id = 50);
```

РЕЗУЛЬТАТЫ СРАВНЕНИЯ



Трино и StarRocks выглядят очень неплохо.

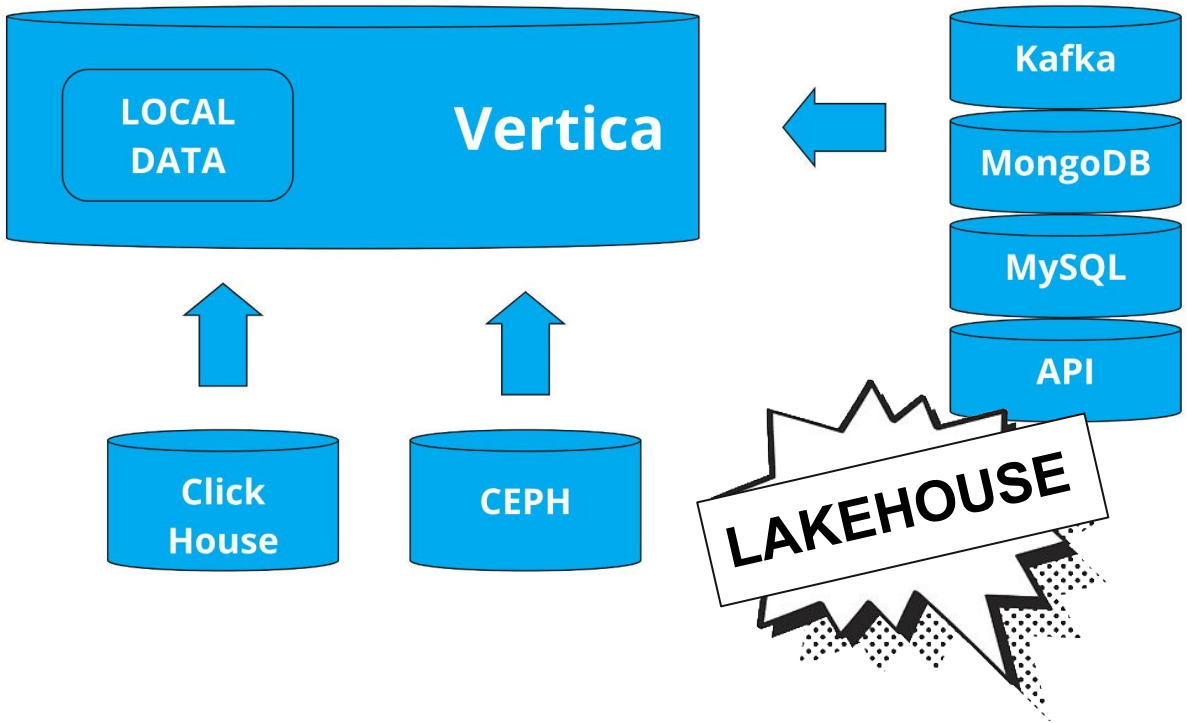
АЛЬТЕРНАТИВНЫЕ РЕЗУЛЬТАТЫ



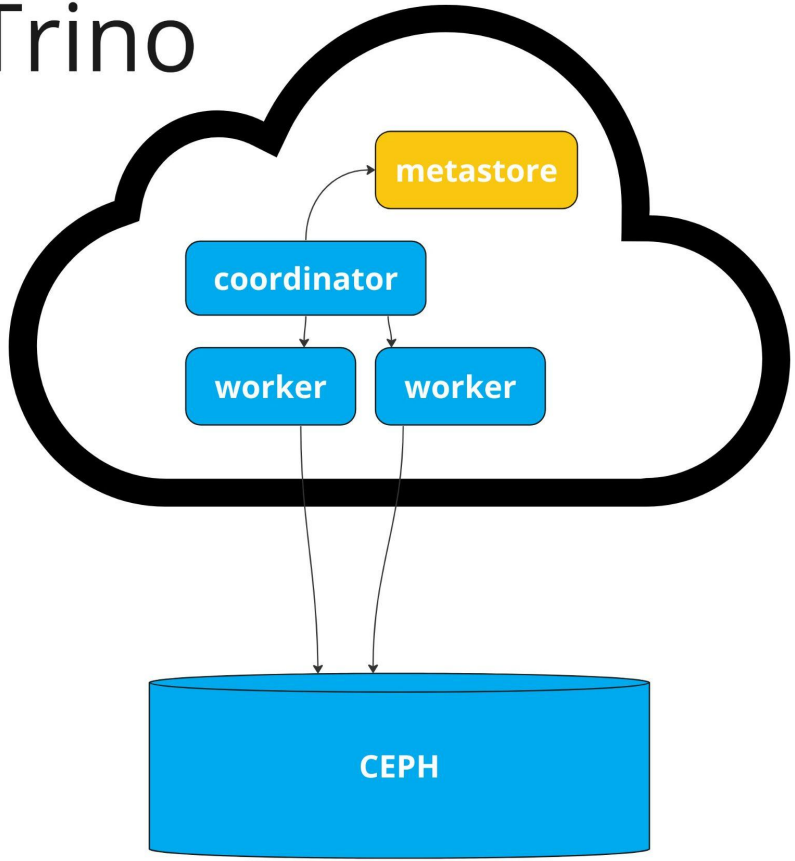
* Choosing A Cloud DBMS: Architectures and Tradeoffs

«Trino хорош и совсем не уступает в производительности»

ПОЧЕМУ НАС ЭТО НЕ ПУГАЕТ



Trino



1
Что мы смотрели

2

3
Так ли хорошо получилось

С какими

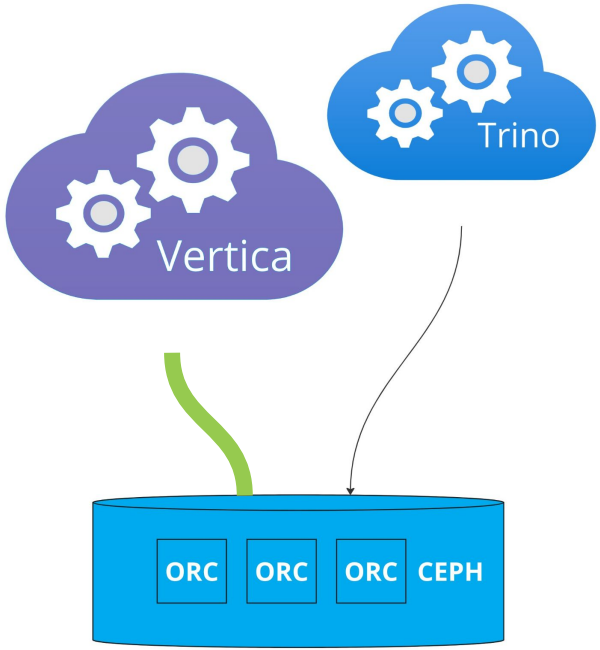
ПРОБЛЕМАМИ СТОЛКНУЛИСЬ

- Как выгружать
- Как переписать
- Как оптимизировать
- Как не потерять

Как вынуть данные

Vertica поддерживает ORC, но

```
select min(event_date)
from arch.click_stream

-- 4500% CPU
-- 10min
```

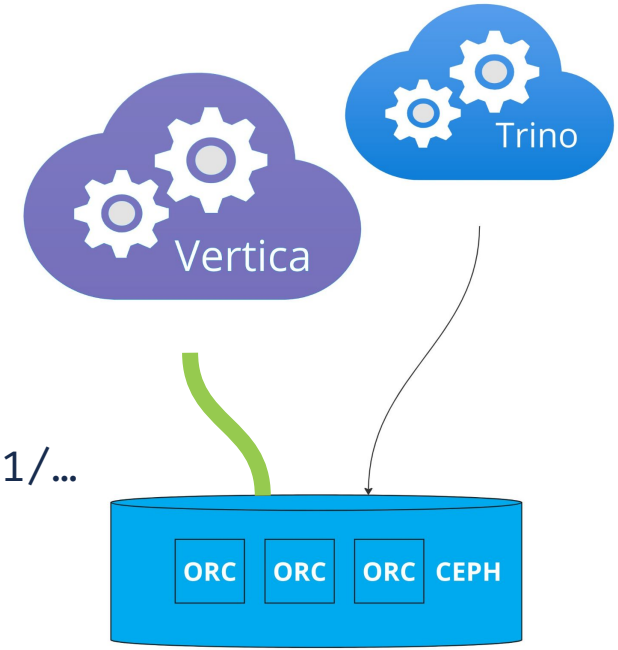


Как вынуть данные

Vertica поддерживает ORC, но

```
PARTITION BY
date_trunc('MONTH', event_date)
```

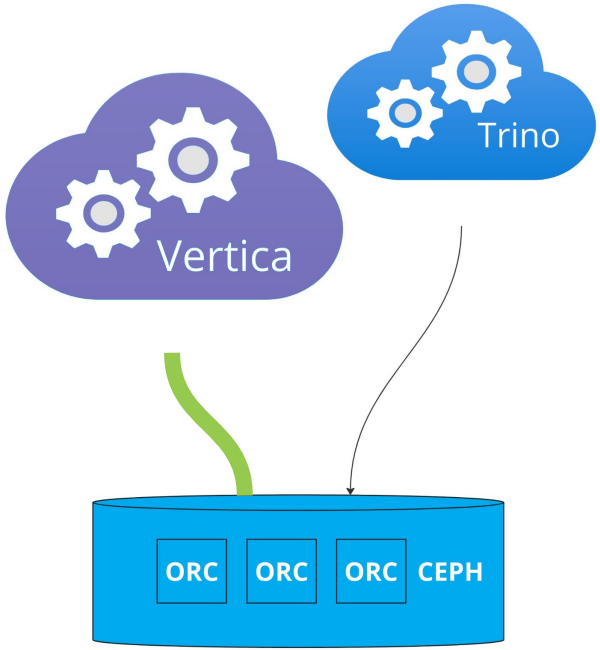
```
≠ hive-partitioning
table_name/event_month=2023-01-01/…
```



Как вынуть данные

Vertica поддерживает ORC, но

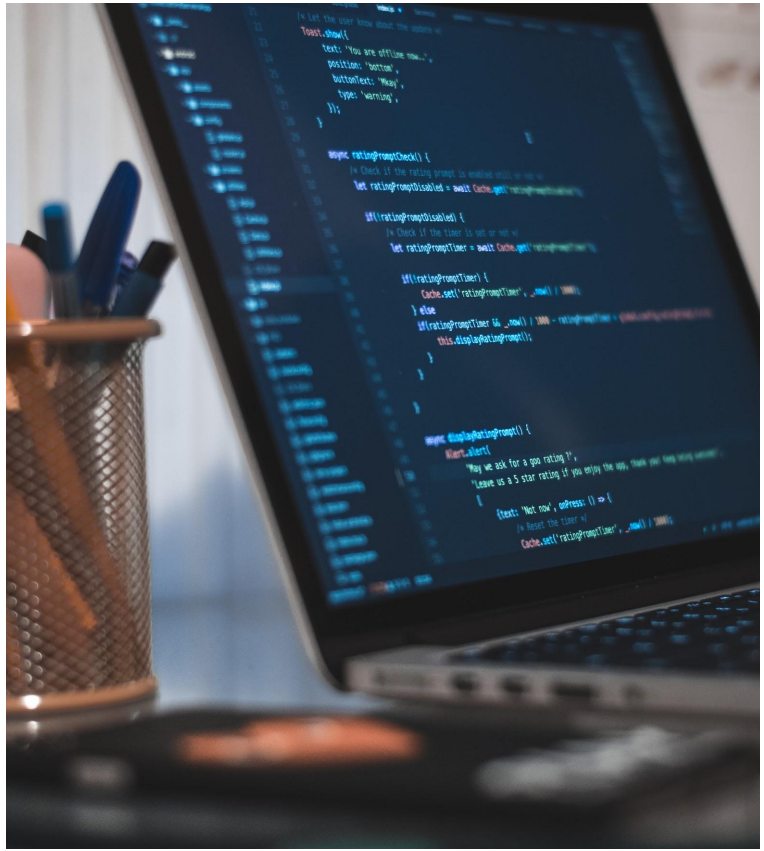
```
WHERE event_date = '2023-01-02'
  AND is_human
  AND type_id in (301,303)
```



Время писать код

Пару спринтов спустя:

- Аналитики могут пользоваться орками без ущерба вертике.
- Получаем контроль над выгрузкой, он нам понадобится в будущем.



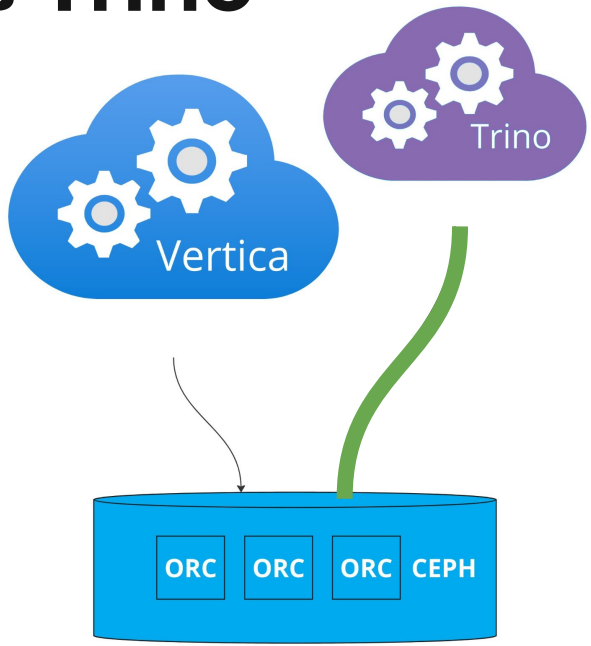
Как получить данные в Trino

Регулярно выгружать из вертики в CEPH:

- создать external-таблицу;
- записать в метастор и собрать статистику.

Считать напрямую в Trino:

- метаданные сами обновляются;
- статистика собирается автоматически.



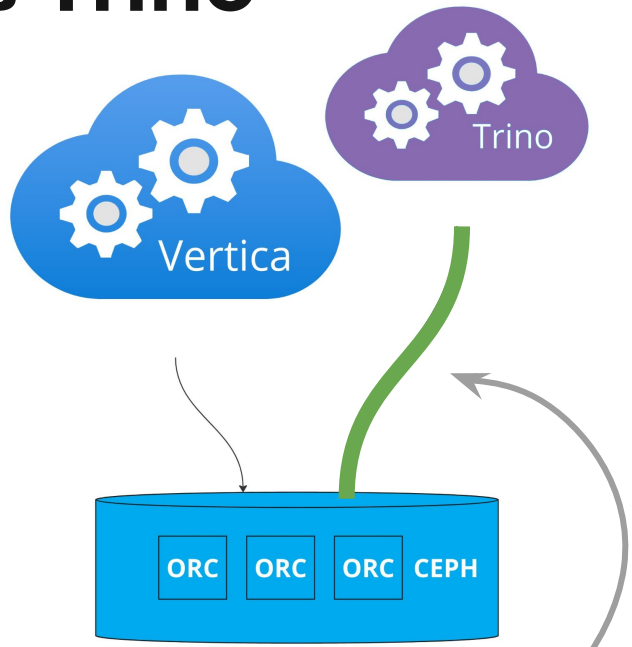
Как получить данные в Trino

Регулярно выгружать из вертики в CEPH:

- создать external-таблицу;
- записать в метастор и собрать статистику.

Считать напрямую в Trino:

- метаданные сами обновляются;
- статистика собирается автоматически.




```
call system.sync_partition_metadata('dma', 'item_day', 'ADD');
analyze dma.item_day with (columns=array['date']);
```

Как транслировать SQL

- Types,
  ANSI casts,
  USING,
  Functions.

```
                vertica │ trino
────────────────────────┼──────────────────────────────────────
                    int │ bigint
                        +
              timestamp │ timestamp(6)
                        +
         SELECT dt::date │ SELECT cast(dt as date)
                        +
WHERE d = '2023-08-01'  │ WHERE d = cast('2023-08-01' as date)
```

Как транслировать SQL

- Trino is designed
  to be a processing engine,
  it doesn't have capabilities
  to store data.

```
create local temp table item_views_users
on commit preserve rows
as select …
segmented by hash(user_id) all nodes;
```

Как транслировать SQL

- Trino is designed
  to be a processing engine,
  it doesn't have capabilities
  to store data.

```
create local temp table item_views_users
on commit preserve rows
as select …
segmented by hash(user_id) all nodes;
```

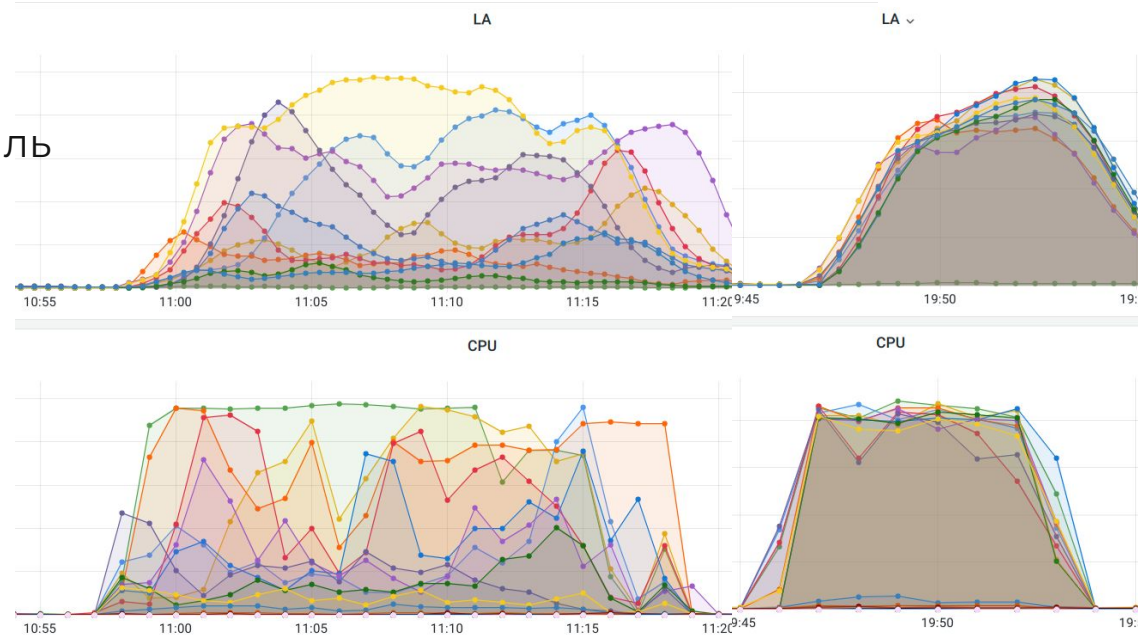
```
create schema dwh.temp_schema_vpupkin
with (location = 's3a://temp/vpupkin');

use dwh.temp_schema_vpupkin;
```

```
create table item_views_users
with (
    bucketed_by = array['user_id'],
    bucket_count = 50
)
as select …
```

Как разогнать расчет

Как понять, что
что-то идет не так:

- неравномерный профиль нагрузки на сервера;
- залипание на стадии выполнения.



Как разогнать расчет

Неэффективный решафлинг:

Отстутствие статистики мешает cost-based оптимизациям:

- join enumeration по дефолту как есть;
- join distribution по дефолту решафлинг, когда мог быть бродкаст.

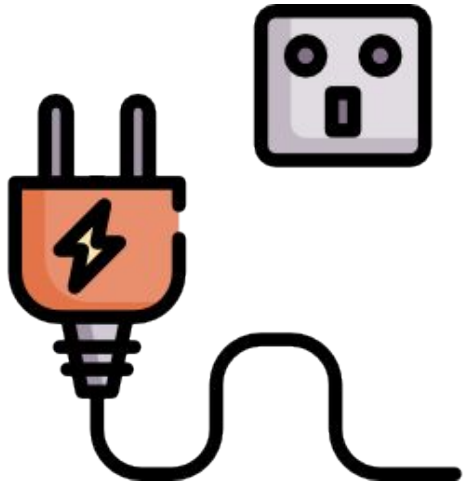
Стандартные буферы для решафлинга смешные:

- `exchange.max-buffer-size` `32MB → 16GB`
- `sink.max-buffer-size` `32MB → 16GB`

Консистентность

Trino использует реализацию транзакций из ORC with (transactional=true) для hive-коннектора.

В OrcLib на C++ для вертики поддержка транзакций не реализована.

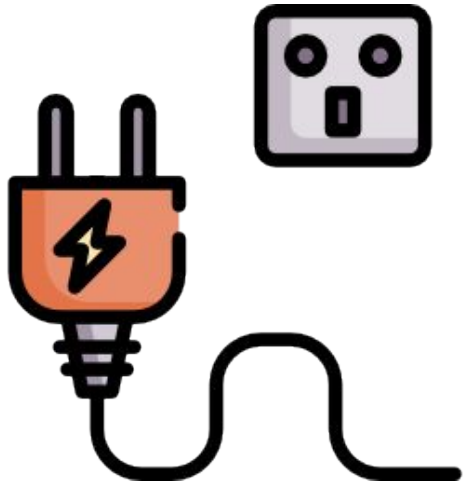


Консистентность

Trino использует реализацию транзакций из ORC with (transactional=true) для hive-коннектора.

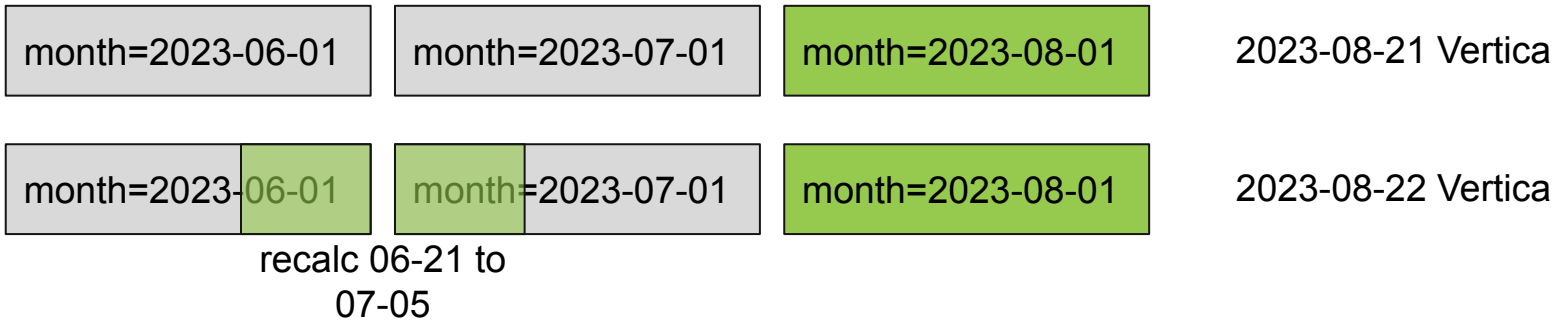
В OrcLib на C++ для вертики поддержка транзакций не реализована:

- поэтому delete только по партициям;
- atomicy и durability на плечах S3 и коннектора.



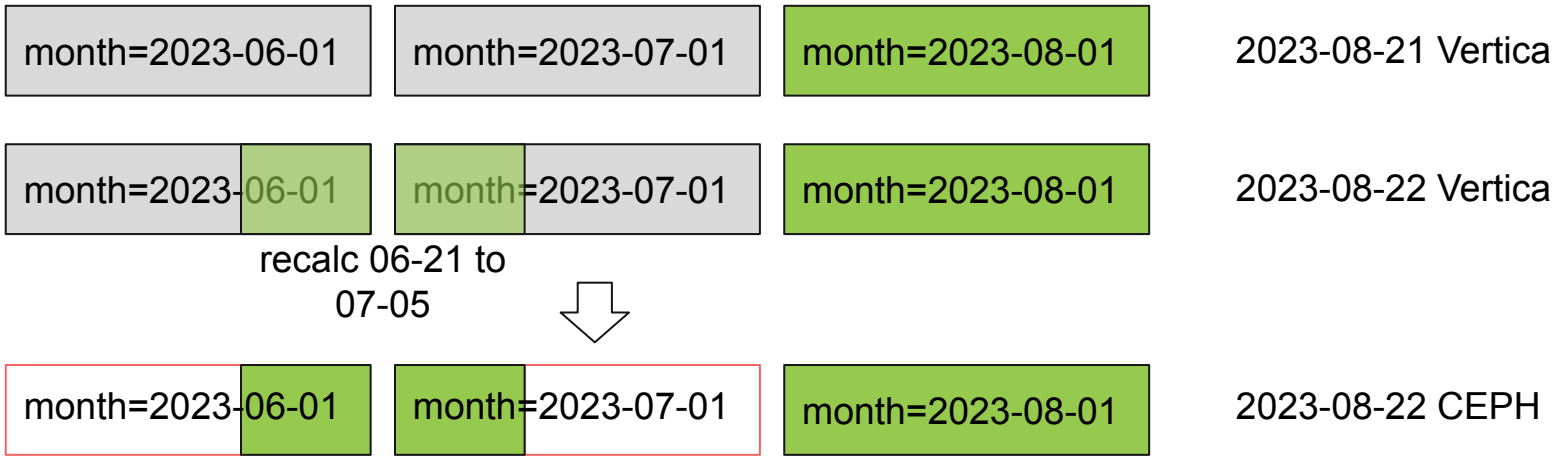
Консистентность

Пересчёты не бьются с партициями в S3.



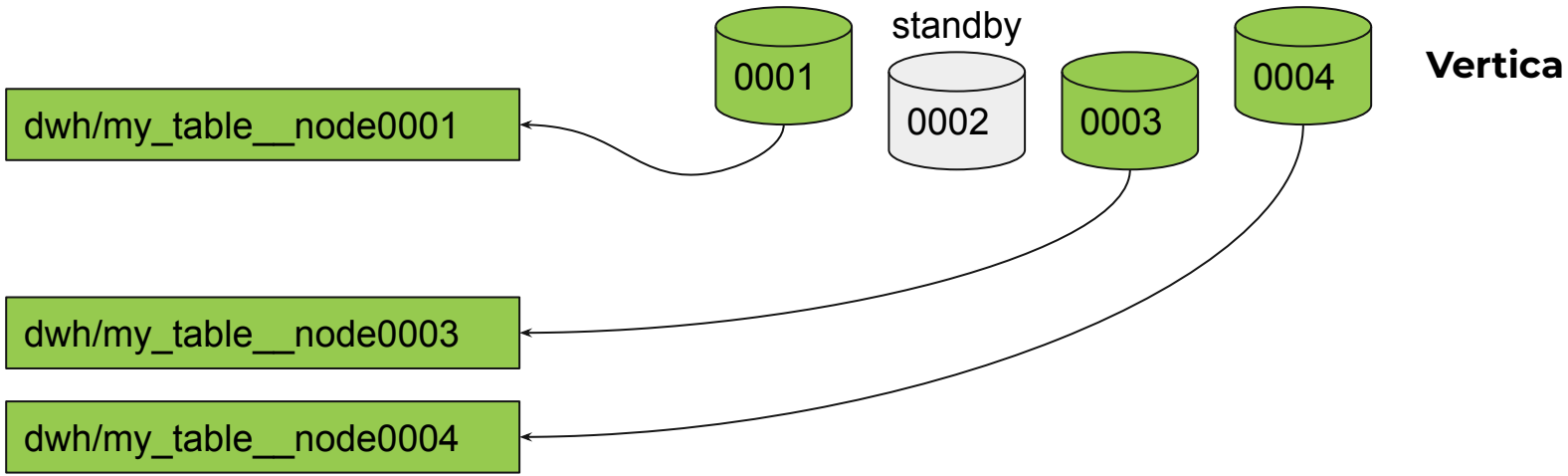
Консистентность

Пересчёты не бьются с партициями в S3,
→ приходится мёрджить файлы на уровне stripe.



Консистентность

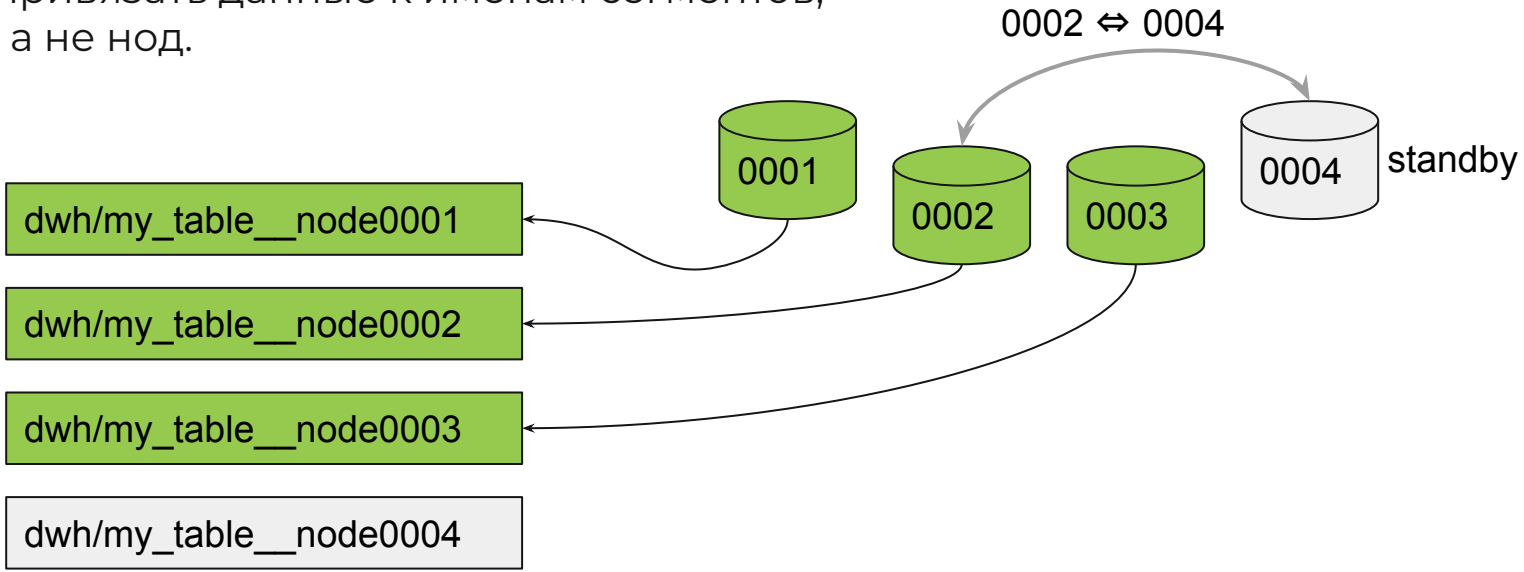
Дубли от подмен нод и расширения кластера.



Консистентность

Дубли от подмен нод и расширения кластера:

→ привязать данные к именам сегментов, а не нод.



1
Что мы смотрели

2
С чем столкнулись

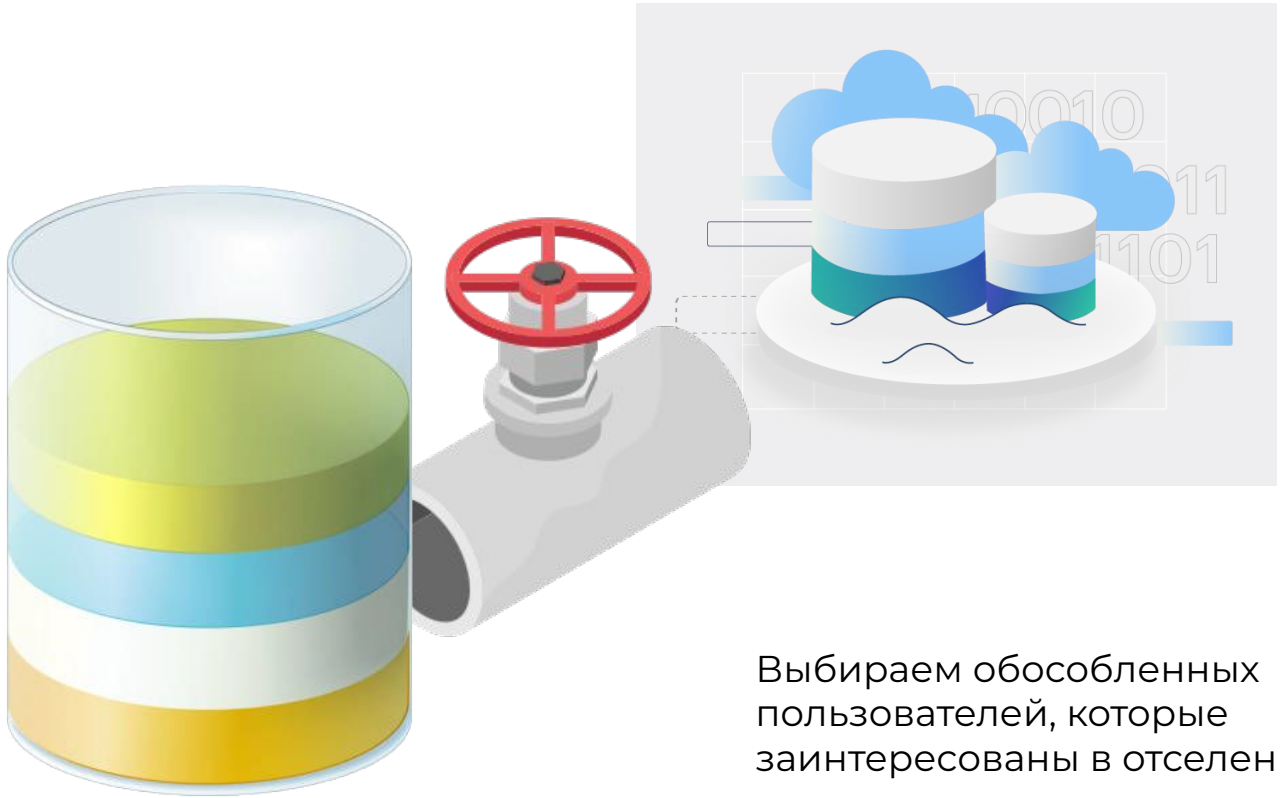
3

ТАК ЛИ ХОРОШО ПОЛУЧИЛОСЬ

- Где мы сейчас
- Устраивает ли скорость
- Достигли ли целей

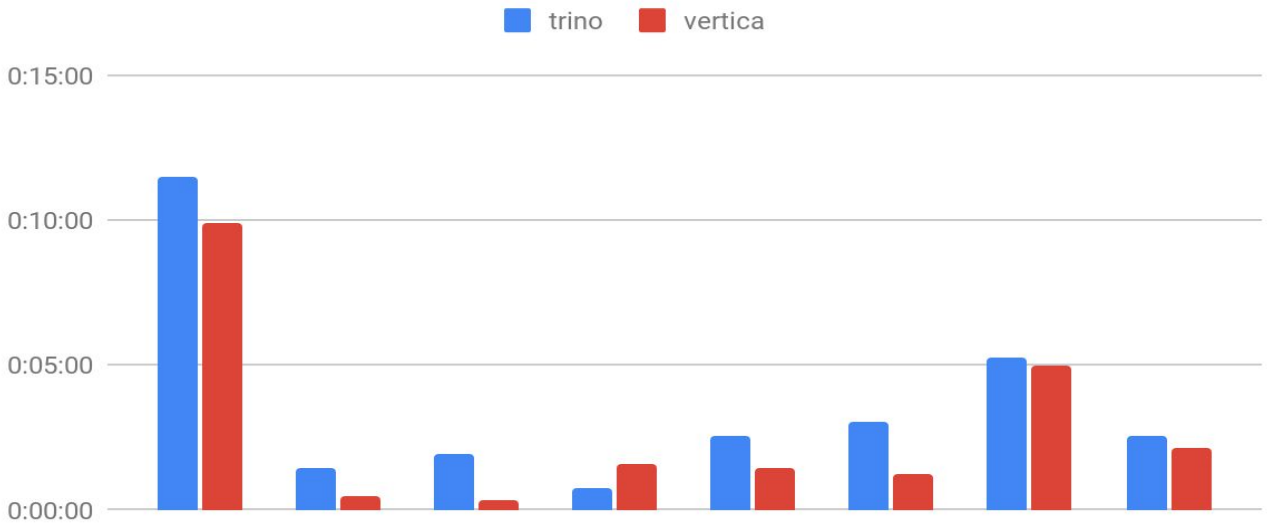
ГДЕ МЫ СЕЙЧАС?

СХЕМА ОТСЕЛЕНИЯ



Выбираем обособленных пользователей, которые заинтересованы в отселении.

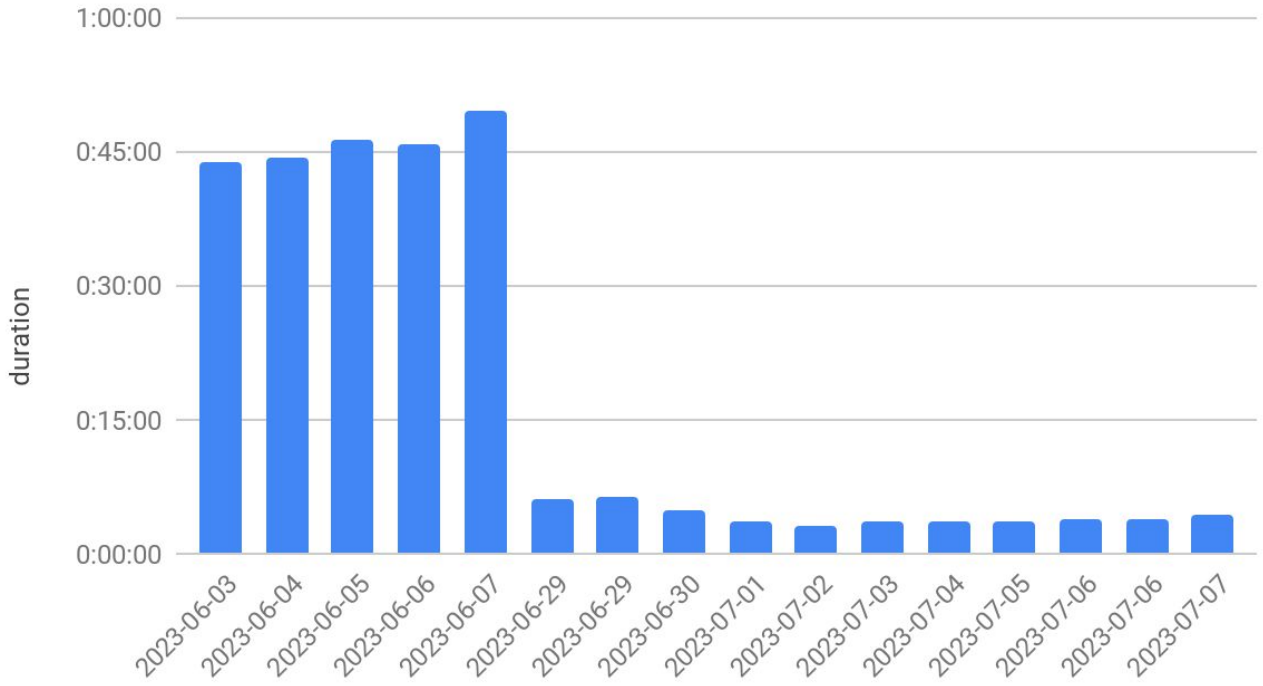
ПРОИЗВОДИТЕЛЬНОСТЬ



Мелкие транслируем asis.

Тяжёлые выравниваем оптимизацией.

МАСШТАБИРОВАНИЕ



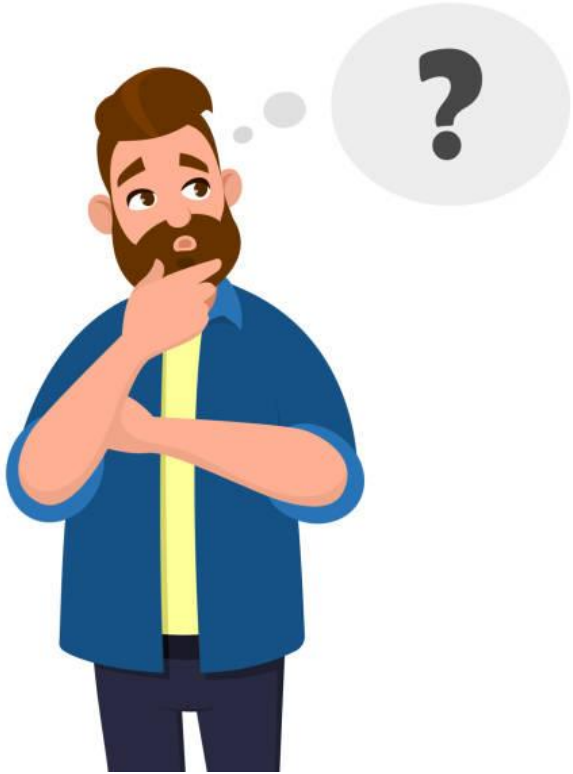
2 недели
vs 5 минут.

3 → 20 nodes:
45 → 5 min

ШУМНЫЕ СОСЕДИ

Вопрос открытый:

- CPU Quotas.
- Kubernetes.



СЛУЧАЙНЫЕ ПАДЕНИЯ

Retry policy:

- QUERY.
- TASK.

Time to recovery:

2min vs 40min

TAKEAWAY

Trino хорош: в нём работают те же оптимизации, к которым мы привыкли в вертике, он бесплатный и масштабируемый

StarRocks супер быстрый, и если у вас меньше петабайта данных, на него стоит посмотреть.

avito.tech

Москва — 2023

phil03

github.com/phil-88

asfilatov@avito.ru

Trino хорош - в нем работают те же оптимизации, что в вертике, он бесплатный и масштабируемый

StarRocks супер на объемах меньше PB