

Ключевые принципы построения дата-платформы в современных условиях

Артеми́й Козы́рь



Что я делаю:

- [Analytics Engineering @ Wheely](#)
- [Performance Marketing Analytics](#)
- Blogging: [Data Apps Design](#) and [Habr](#)
- ex DE & DWH @ Sber, Sibur, Moscow Exchange

С чем работаю:

- DWH: Snowflake, Clickhouse, [Redshift](#)
- Transformations: [Data Build Tool](#)
- BI: [Looker](#), Metabase, Superset, Datalens
- AWS, Yandex.Cloud, Cloud.ru
- [SQL](#), Python, Bash, Docker, Github Actions

Ключевая идея #1:

Создавайте ценность

Бизнес-ценность

- Отчетность по KPI
- Проактивная / Предиктивная аналитика
- Продуктовая аналитика

Отчетность по ключевым метрикам компании

- Группы метрик: Finance, Operations, Marketing, Customer Support
- Измерения: время, продукты, гео, устройства, и т.д.
- Продвинутая аналитика: Period over Period, % Change, Rolling metrics, Target comparison, Geospatial
- Ночные / часовые / near real time обновления метрик

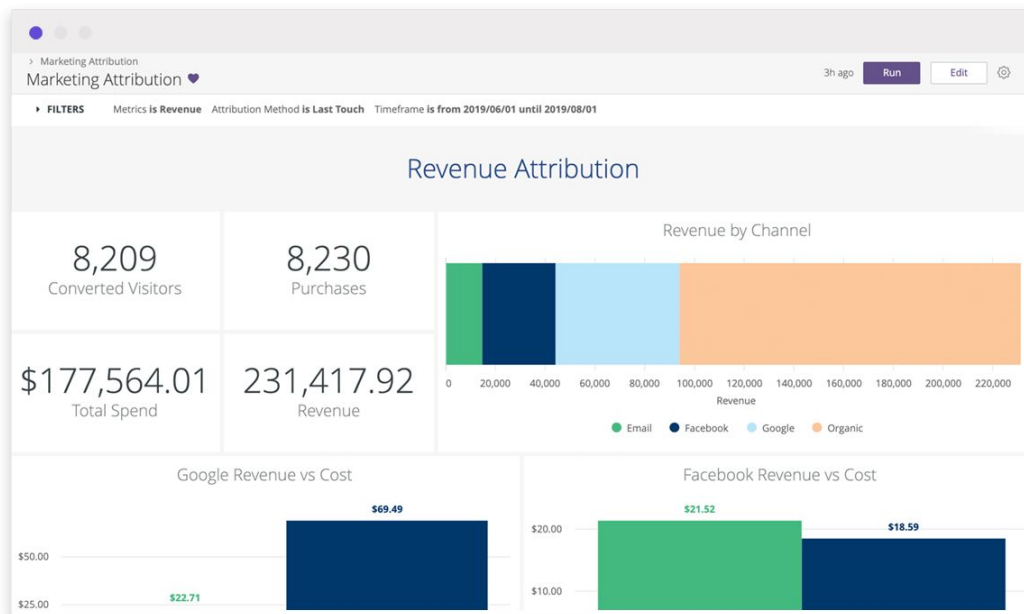
Проактивная / Предиктивная аналитика

- Customer Acquisition: Сквозная аналитика, RFM, сегментация
- Reverse ETL: обогащение операционных систем
- Recommendations / Next best offer
- Churn prediction
- Fraud monitoring

Продуктовая аналитика

- Product metrics: формирование и отслеживание
- Funnel analysis: анализ конверсий и воронок, UX
- Cohort analysis: установки vs. retention
- A/B тестирование инициатив

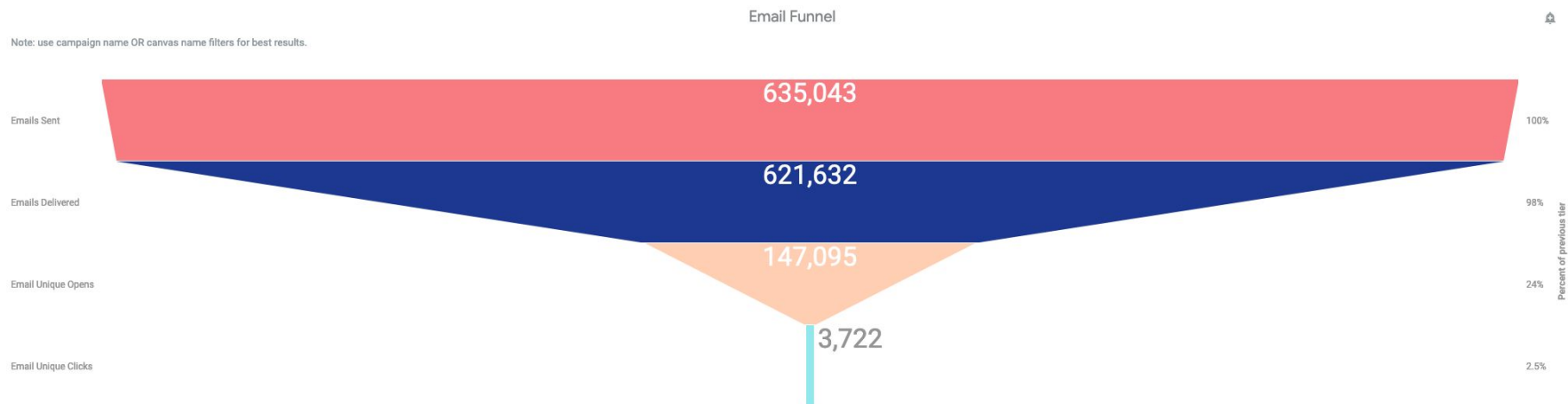
Маркетинг и атрибуция трафика



Когортный анализ

Cohort Period >		0	1	2	3	4	5	6	7	8
Cohort Week ↑ ⚙		% Retention	% Retention	% Retention	% Retention	% Retention	% Retention	% Retention	% Retention	% Retention
1	2021-06-28	⌕	⌕	⌕	⌕	⌕	⌕	⌕	⌕	⌕
2	2021-07-05	100%	95%	80%	73%	68%	56%	63%	61%	56%
3	2021-07-12	100%	88%	73%	78%	78%	68%	60%	68%	63%
4	2021-07-19	100%	84%	82%	58%	53%	58%	63%	61%	53%
5	2021-07-26	100%	89%	77%	73%	75%	66%	66%	59%	61%
6	2021-08-02	100%	85%	78%	61%	75%	61%	64%	61%	54%
7	2021-08-09	100%	83%	74%	72%	67%	59%	54%	61%	54%
8	2021-08-16	100%	90%	82%	77%	72%	69%	62%	62%	49%
Total		100%	86%	75%	71%	68%	61%	59%	58%	54%

Email funnel (conversions)



Ключевая идея #2:

Решайте прикладные задачи

Фокус-направления в работе с данными

- Data Integration
- Data modeling
- Consuming data
- Team work
- Data Quality
- Observability + Documentation

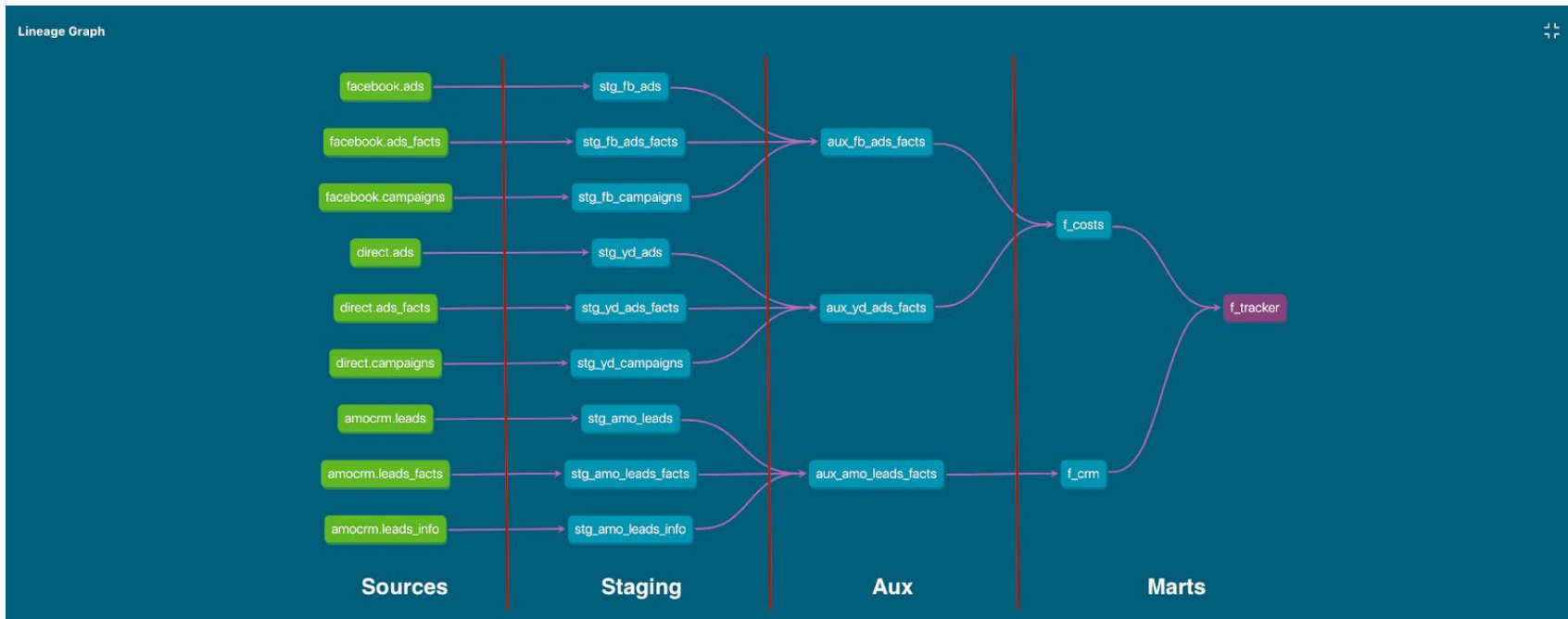
Data Integration

- Собираем данные из систем-источников: CRM, Backend, Marketing, Web trackers
- Делаем это регулярно в инкрементальной манере (дельты) вплоть до real time replication
- Копируем данные источников 1:1 (= принцип ELT)

Data modeling

- Парсинг, маппинг, обогащение, агрегирование данных
- Heavy Lifting в мощной колоночной СУБД (Clickhouse)
- Понятная последовательность операций и слоев преобразований (DAG)
- Всё на высокоуровневом и декларативном SQL (низкий порог входа)
- Результат: слой витрин данных (оптимизированные для чтения таблицы)

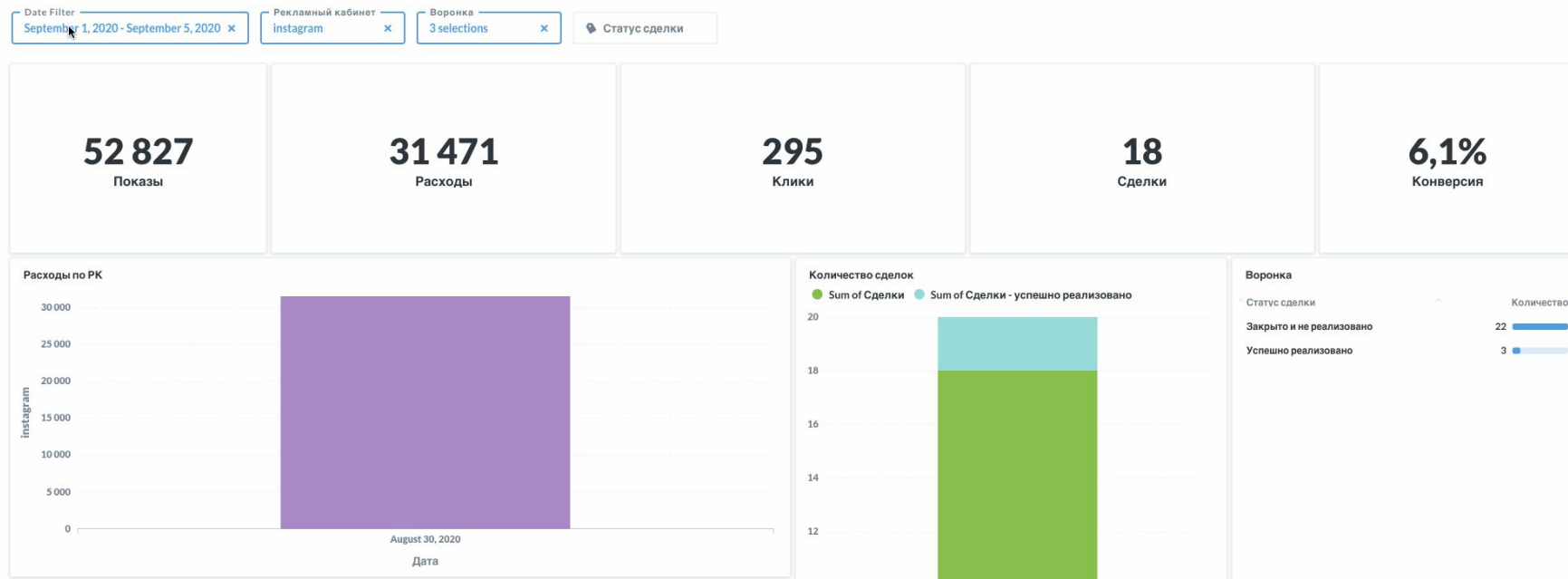
Data modeling (Layered architecture)



Consuming data

- Набор готовых Dashboards для бизнеса
- Конструктор для сбора новых Dashboards (Self-service)
- API-доступ к витринам для других команд
- Time 2 Market: изменения (фичи, фиксы) в течение рабочего дня

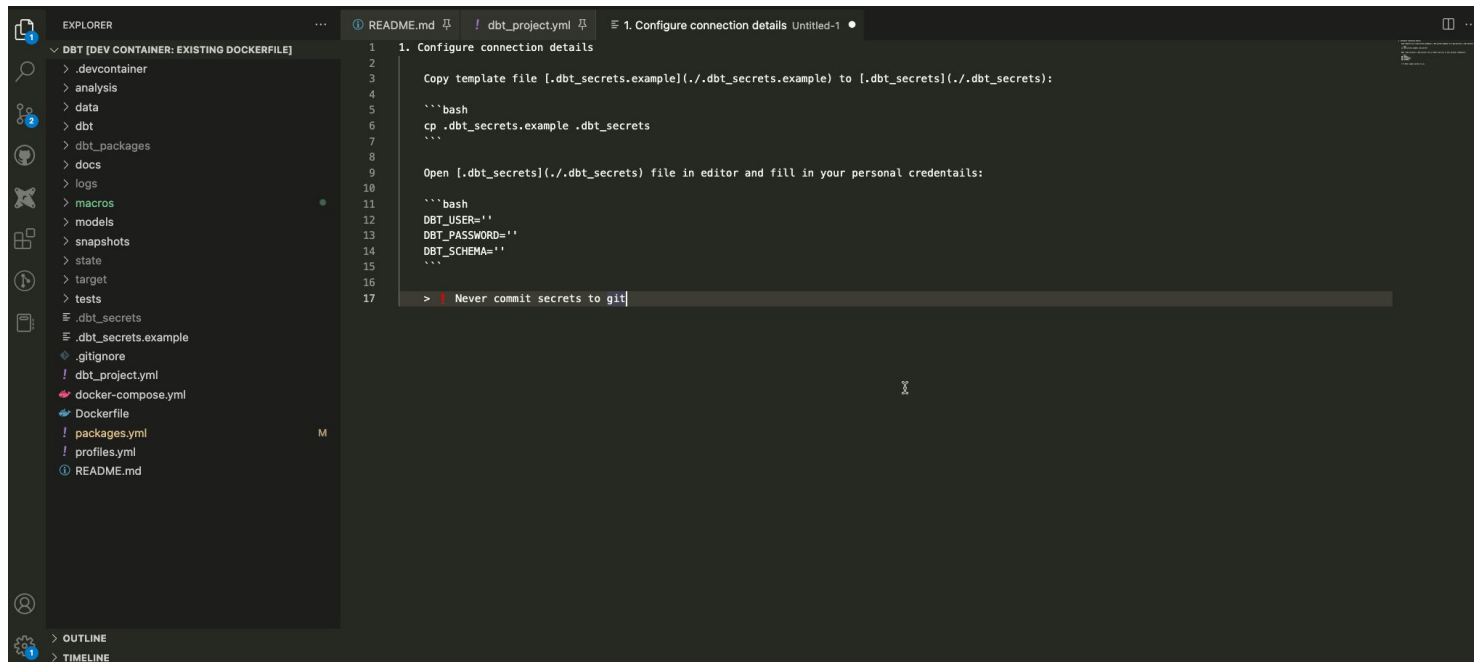
Consuming data (Metabase)



Team work

- Инструменты для командной работы над сложными задачами
- Всё версионруется в git: история изменений, быстрый rollback, использование branches
- Готовые среды разработки (devcontainers) и песочницы (sandbox)
- Работа с кодом через Pull Requests & Peer Reviews
- Набор автоматизированных проверок (тестов) перед публикацией в Prod

Team work (Instant Dev environment)



Team work (Github collaboration)

March 28, 2021 – April 28, 2021

Period: 1 month ▾

Overview

45 Active Pull Requests

0 Active Issues

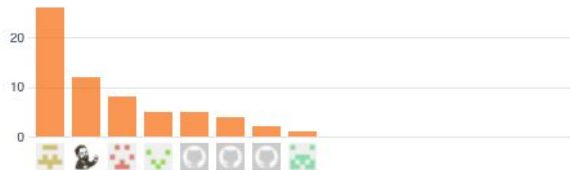
44
Merged Pull Requests

1
Open Pull Request

0
Closed Issues

0
New Issues

Excluding merges, **8 authors** have pushed **44 commits** to master and **50 commits** to all branches. On master, **105 files** have changed and there have been **2,955 additions** and **1,826 deletions**.

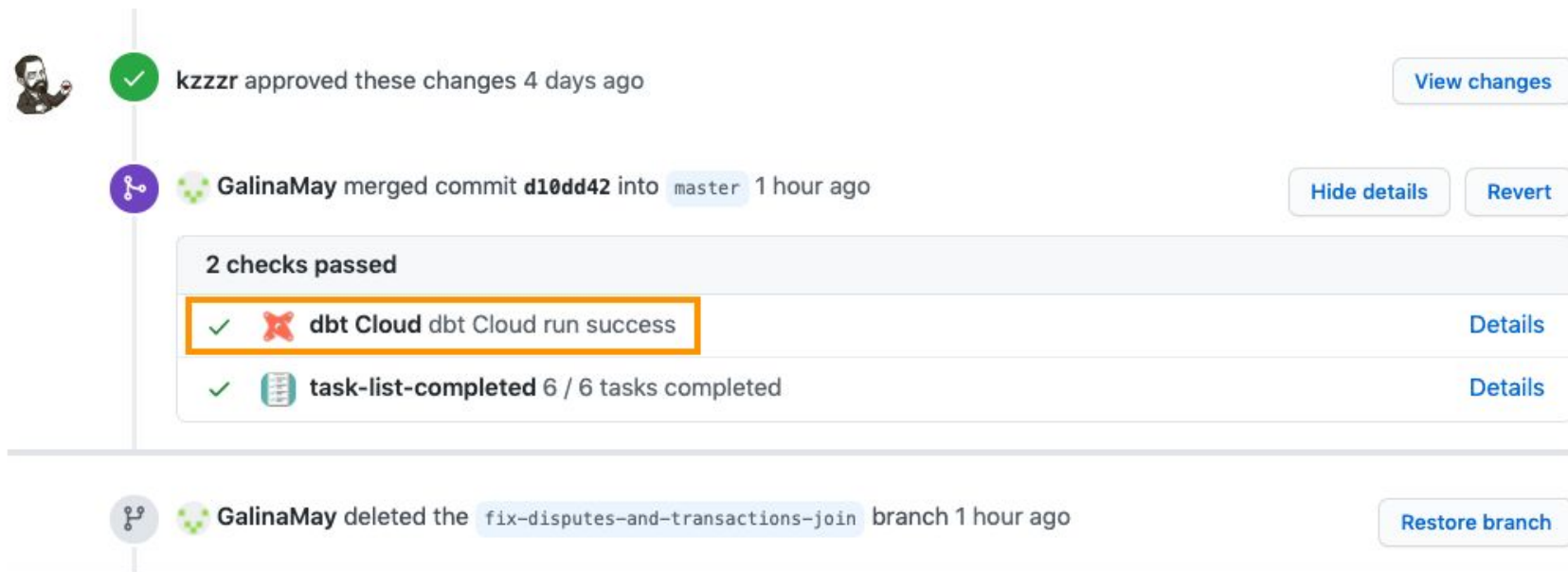


44 Pull requests merged by 5 people

Data Quality

- Базовые тесты наборов данных: not null, unique, foreign keys
- Source freshness tests (все ли интеграции в порядке?)
- Generic tests: проверка бизнес-логики и ожиданий

Data Quality (Automated checks)



The screenshot displays a GitHub commit history with three entries. The first entry shows a user approval. The second entry, which is the focus, shows a merge of commit `d10dd42` into the `master` branch, accompanied by a summary of two passed checks. The first check, `dbt Cloud dbt Cloud run success`, is highlighted with an orange border. The second check is `task-list-completed 6 / 6 tasks completed`. The third entry shows a branch deletion. Each entry includes a 'View changes' button, while the merge entry also includes 'Hide details' and 'Revert' buttons, and the branch deletion entry includes a 'Restore branch' button.

  kzzzr approved these changes 4 days ago [View changes](#)

  GalinaMay merged commit `d10dd42` into `master` 1 hour ago [Hide details](#) [Revert](#)

2 checks passed

-   **dbt Cloud** dbt Cloud run success [Details](#)
-   **task-list-completed** 6 / 6 tasks completed [Details](#)

  GalinaMay deleted the `fix-disputes-and-transactions-join` branch 1 hour ago [Restore branch](#)

Data Quality (Source freshness)

Data Sources ⓘ						
Warning SOME DATA SOURCES REQUIRE ATTENTION		39 OK	1 WARNING	0 STALE	an hour ago LAST SNAPSHOT	PROD - SOURCE FRESHNESS > JOB
				Run #16011012 > RUN		
SOURCE	LOADER / LOADED	LAST SNAPSHOT		STATUS		
> app_documents_prod (1)	Hevo	an hour ago		OK		
> communication_prod (4)	Hevo	an hour ago		OK		
> dq (10)	Airflow			UNTRACKED		
> external (2)	S3	an hour ago		OK		
> financial_service (6)	Hevo	an hour ago		OK		
> flights_prod (2)	Hevo	an hour ago		OK		
> gs (5)	Hevo			UNTRACKED		
> manual_load (7)	Manual			UNTRACKED		
> marketing_external_api (3)	analytics_etl (ECS)	an hour ago		WARNING ●		
> marketing_hevo (2)	hevo			UNTRACKED		

Observability + Documentation

- Самый удобный способ - вебсайт
- Доступность документации и графов зависимостей (DAG / Lineage)
- Простой поиск источников, витрин, метрик и методов их расчета
- Документация там же, где и код

Observability + Documentation

- [Live demo](#)

Ключевая идея #3:

Используйте современные решения

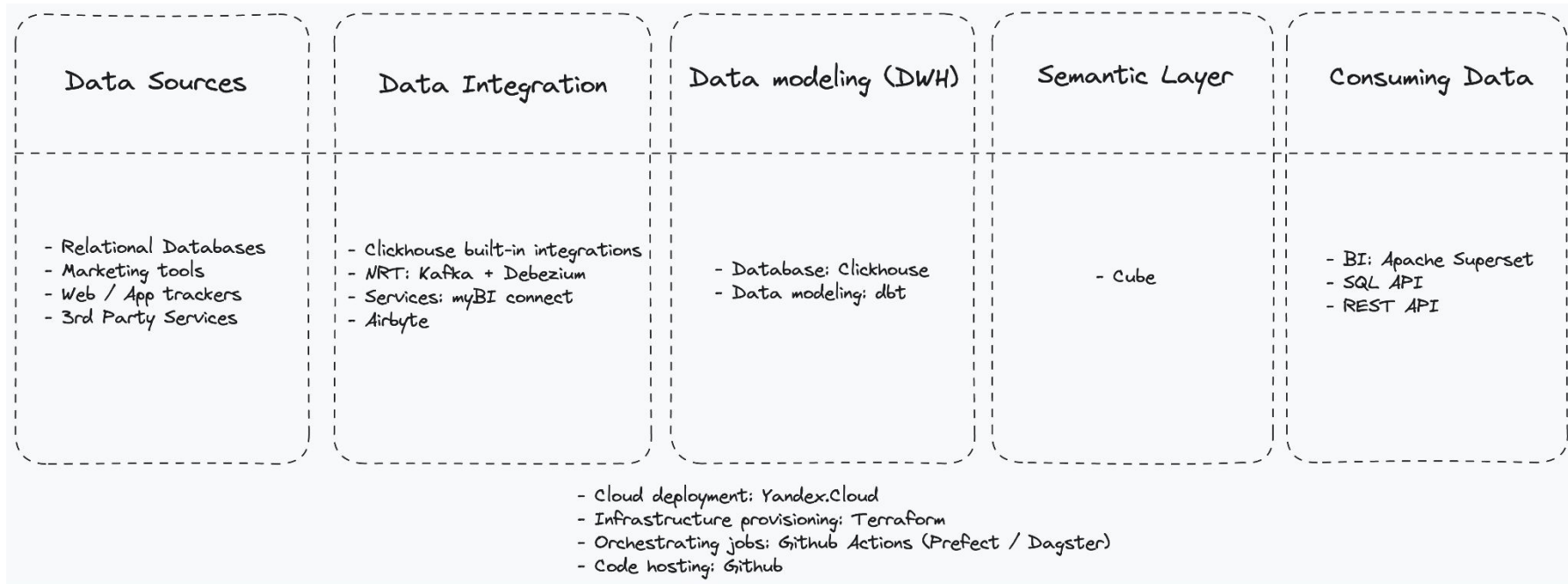
Modern Data Stack

- Современные и передовые решения
- Фокус на создании бизнес-ценности
- Простота администрирования, сопровождения, исправления ошибок

Modern Data Stack (пример)

- Infrastructure as Code (Terraform)
- Clickhouse
- dbt for T-transformations
- Cube for semantic layer
- Superset / Datalens for BI
- Github for code collaboration

Sample architecture view



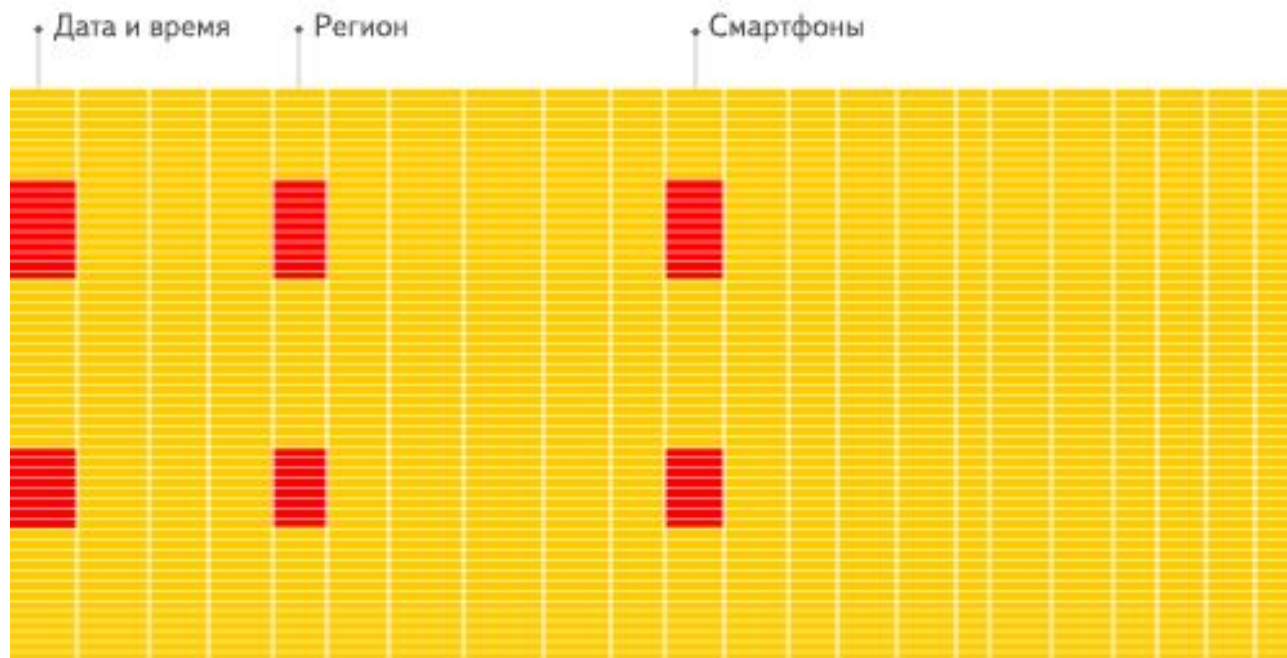
Infrastructure as Code (Terraform)

- Подход IaC = Инфраструктура как код
- Развертывание, модифицирование, удаление ресурсов за минуты
- Прозрачность трат
- Ролевая модель Identity Access Management (IAM)

Clickhouse

- Одна из лучших в мире СУБД для аналитики
- Sub-second queries
- Не нуждается в представлении

Clickhouse



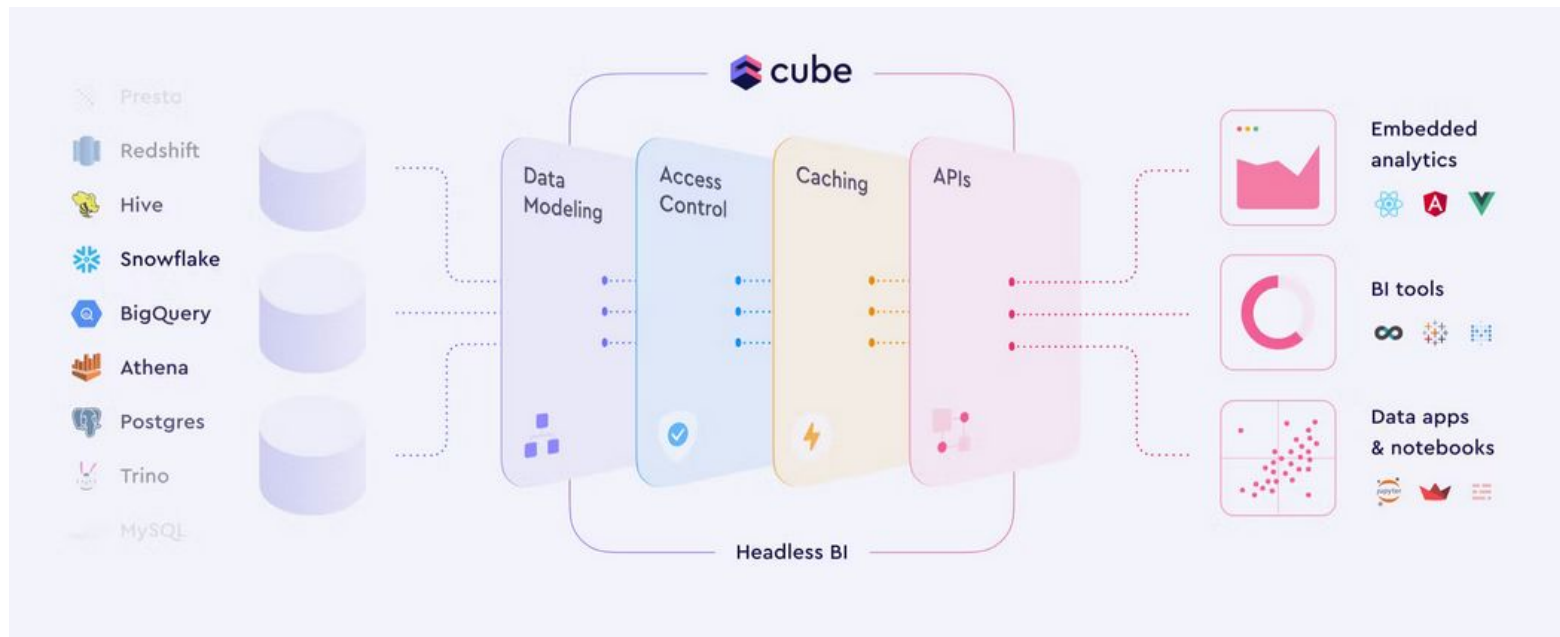
dbt for T-transformations

- SQL на стероидах
- Версионирование, шаблонизация кода
- Макросы, модули, адаптеры

Cube for semantic layer

- Единый семантический слой - без привязки к BI tools
- Задаем правила расчета бизнес-метрик (yaml)
- Формируем связи между витринами (факт - измерение)
- Cube генерирует правильные запросы к СУБД

Cube for semantic layer



Github for code collaboration

- Хостинг кода
- Github Actions for jobs orchestration
- Codespaces для Dev-сред (даже в браузере)

Github Actions

The screenshot displays the Github Actions interface for a workflow named 'Clickhouse CI'. On the left sidebar, the 'Jobs' section is active, showing a list of jobs: 'Clickhouse CI' (selected), 'Run details', 'Usage', and 'Workflow file'. The main panel shows the 'Clickhouse CI' job, which succeeded 2 days ago in 2m 41s. The job log is expanded for the 'Run dbt debug --target ci' step, showing a successful connection to Clickhouse with various configuration details. Below the log, a list of steps is shown, including 'Run dbt deps --target ci', 'Run dbt run-operation init_s3_sources --target ci', 'Run dbt build --target ci', 'Post Run actions/checkout@v3', 'Stop containers', and 'Complete job'.

Clickhouse CI
succeeded 2 days ago in 2m 41s

Search logs

Run dbt debug --target ci 2s

```
18
19 Connection:
20 driver: None
21 host: clickhouse
22 port: 8123
23 user: default
24 schema: default
25 retries: 1
26 database_engine: None
27 cluster_mode: False
28 secure: False
29 verify: True
30 connect_timeout: 10
31 send_receive_timeout: 300
32 sync_request_timeout: 5
33 compress_block_size: 1048576
34 compression:
35 check_exchange: True
36 custom_settings: None
37 use_lw_deletes: False
38 Connection test: [OK connection ok]
39
40 All checks passed!
```

> Run dbt deps --target ci 3s

> Run dbt run-operation init_s3_sources --target ci 5s

> Run dbt build --target ci 1m 37s

> Post Run actions/checkout@v3 0s

> Stop containers 1s

> Complete job 0s

Спасибо. Вопросы?

