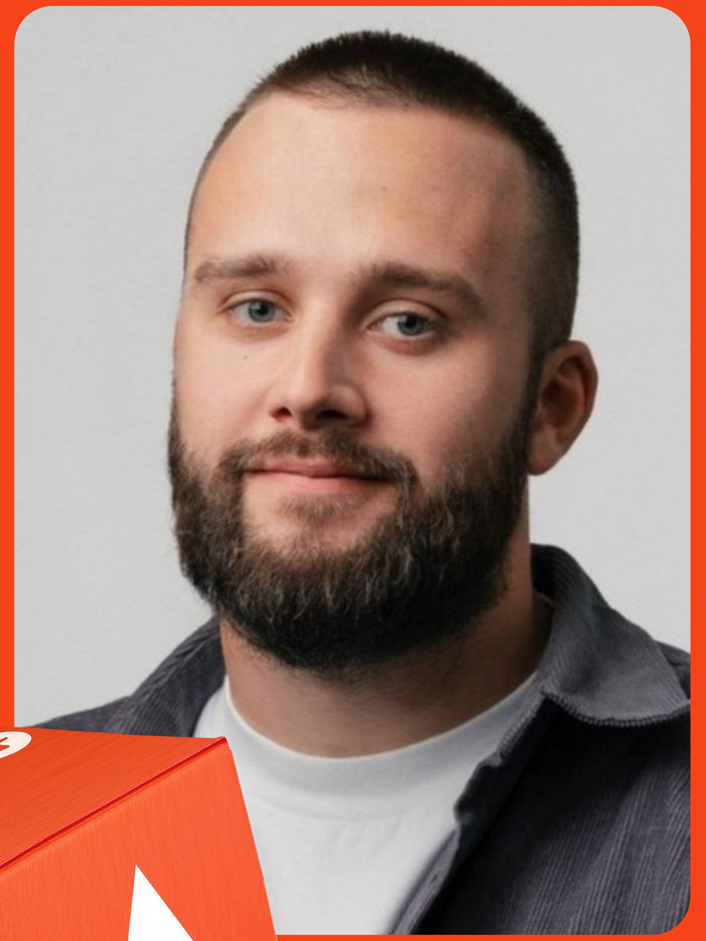
критерии хорошей платформы данных



Влад Гоцуляк
Руководитель платформы данных в Яндекс Доставке

ГОРОДСКИЕ СЕРВИСЫ

познакомимся!

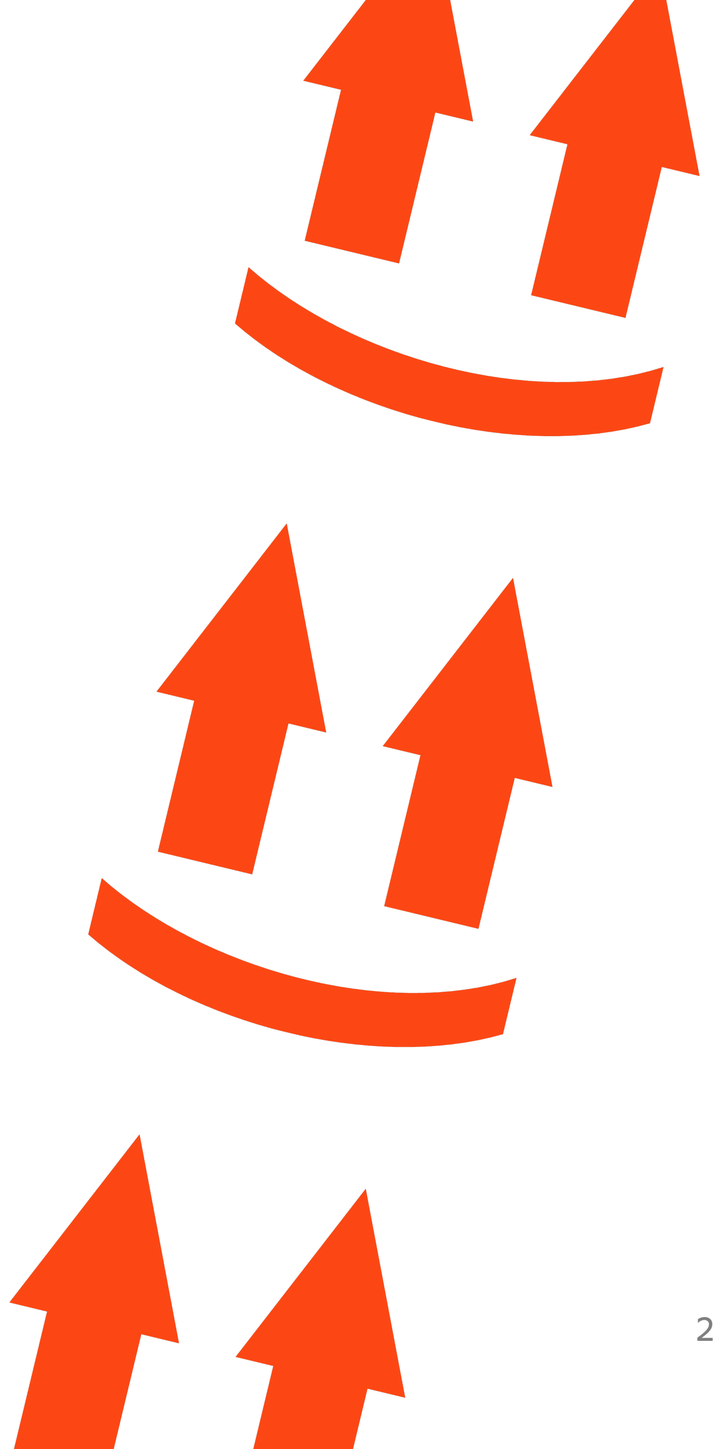
01 Выпускник физтеха

02 Дата-инженерил в прошлом,
теперь менеджер большой команды

03 Руковожу платформой данных в Яндекс Доставке

04 До этого работал CDO Eapteka, MLE в X5, DE в Sber

05 Люблю кайтсерфинг, сноуборд и вейкборд



о чём сегодня поговорим

критерии хорошей платформы данных

01 Стабильность инфраструктуры и платформы

02 Качество данных

03 Скорость отчётов и аналитических запросов

04 Полнота покрытия бизнесовых метрик

05 Высокая частотность использования бизнесом и аналитикой

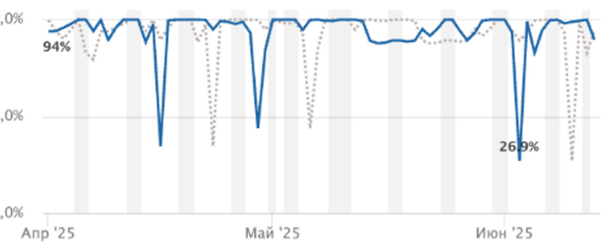
06 Эффективность использования железа

07 Инженерная культура аналитиков

frastructure Index
ows main infrastructure systems uptime

3.9%
al for period

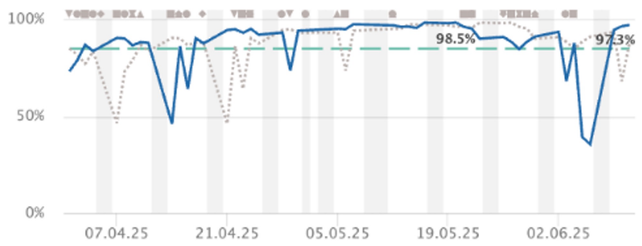
Infrastructure



Stability Index
shows how well our object covered by SLA checks, SLA and Tasks uptime

87.2%
total for period

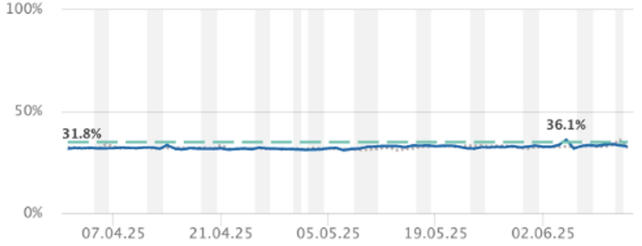
Stability



Quality Index
shows how well our object covered by DQ checks and how well this checks executed

32.4%
total for period

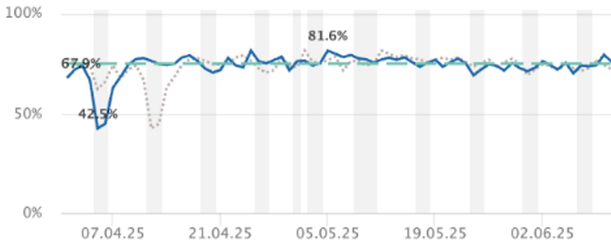
Quality



Query Performance Index
shows how efficient and fast SQL-queries executed

74.8%
total for period

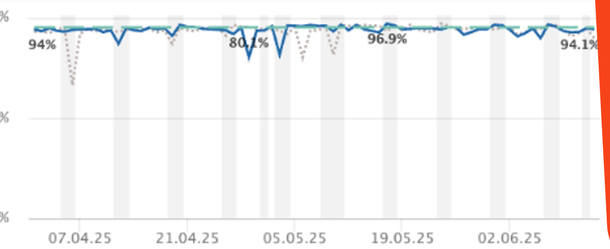
Query Performance



ashboard Performance Index
ows how stable and fast BI dashboards work

3.7%
al for period

Dash Performance

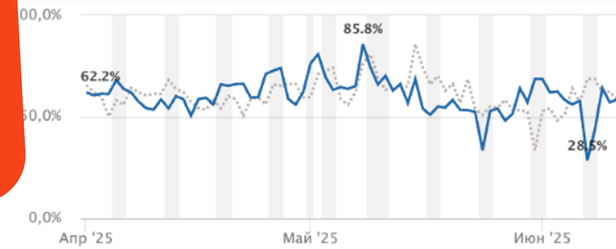


Metric Hierarchy Index
shows the coverage of business-metrics

Dashboard Usage Index
shows the usage of data platform BI dashboards vs. all dashboards in 05Delivery

59.7%
total for period

Dashboard Usage

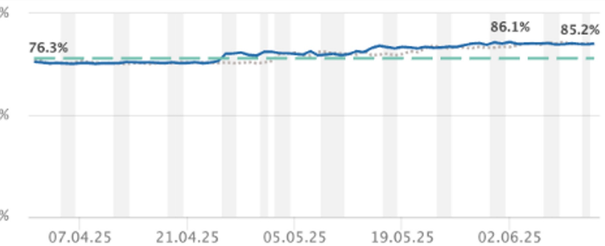


сделали инструмент для контроля метрик и процессов

ngineering Culture Index
ows how well analysts follow rules when creating Nirvana pipelines

0.5%
al for period

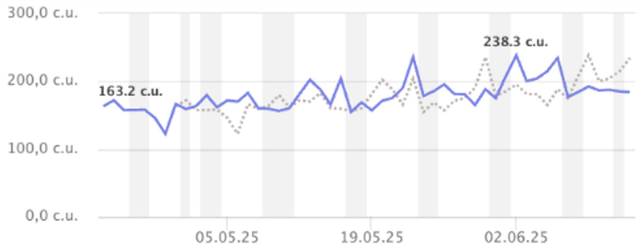
Engineering Culture



Efficiency Index
shows the approximate quota costs

9314 c.u.
total for period

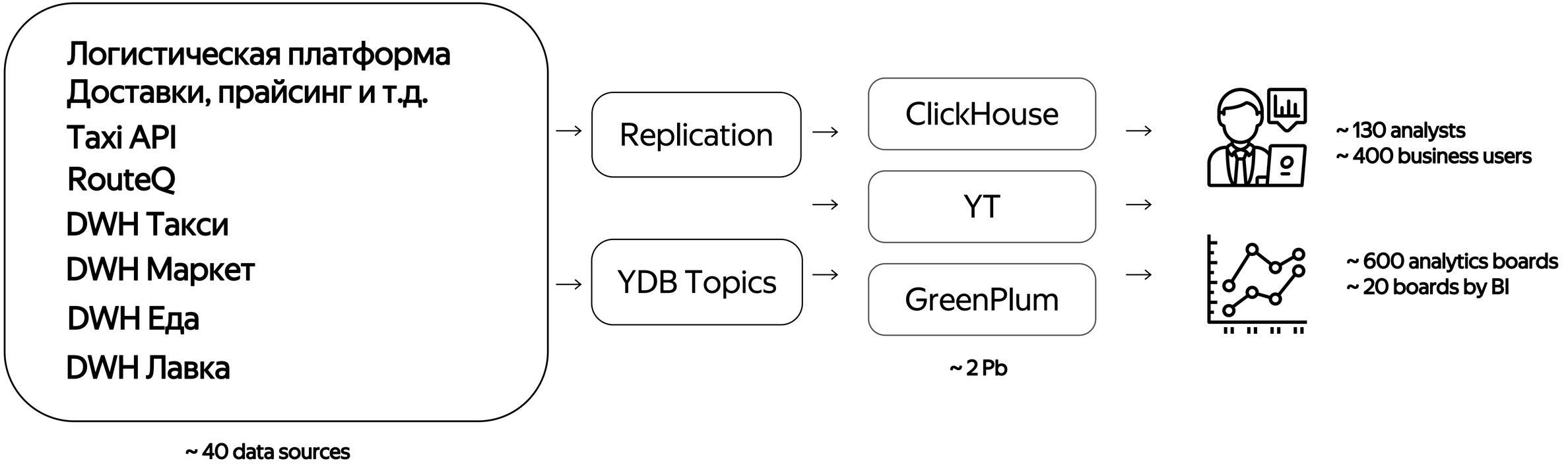
Efficiency



Indexes by Domains
summary by domains, shows total values for selected period

DOMAIN	Stability Index	Quality Index	Query Performance Index	Dashboard Perfomance Index	Metrics Hierarchy Index	Tables Usage Index	Dashboard Usage Index	Engineering Culture Index	Efficiency Index c.u.
supply	89,2%	26,4%	61,7%	63,6%	66,1%	42,0%	66,6%	81,3%	739,4 c.u.
demand	88,1%	39,0%	67,7%	63,9%	32,0%	48,4%	18,9%	89,5%	1 817,6 c.u.
finance	87,9%	47,8%	81,3%	82,6%	76,6%	48,7%	56,0%	68,8%	638,4 c.u.
efficiency	87,8%	19,7%	76,2%	62,8%	53,8%	35,9%	18,8%	71,0%	3 490,4 c.u.
support	60,0%	31,0%	85,4%	76,5%	0,0%	44,7%	0,0%	*null*	1 180,5 c.u.
bets	80,9%	51,0%	71,4%	68,8%	3,6%	46,6%	0,0%	75,8%	1 429,6 c.u.
core	88,5%	26,9%	*null*	92,2%	0	*null*	98,7%	90,2%	18,7 c.u.

внедрили для архитектуры всей платформы Яндекс Доставки



01

метрики стабильности инфраструктуры

следим за стабильностью нашей инфраструктуры через Infrastructure Index

Infrastructure Index

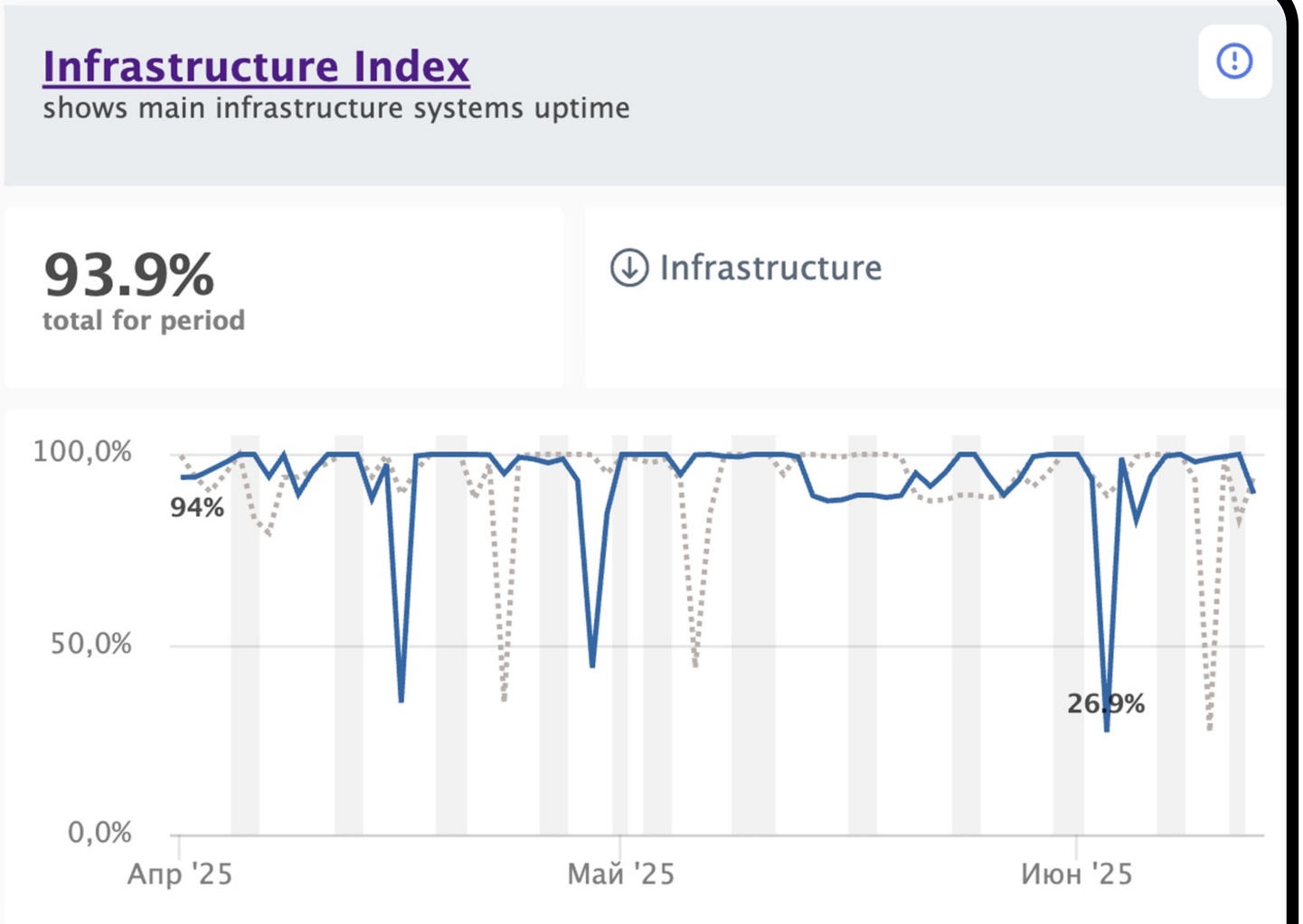
This metric shows the uptime of main infrastructure systems.
Index calculated as average system uptimes.

Formula:

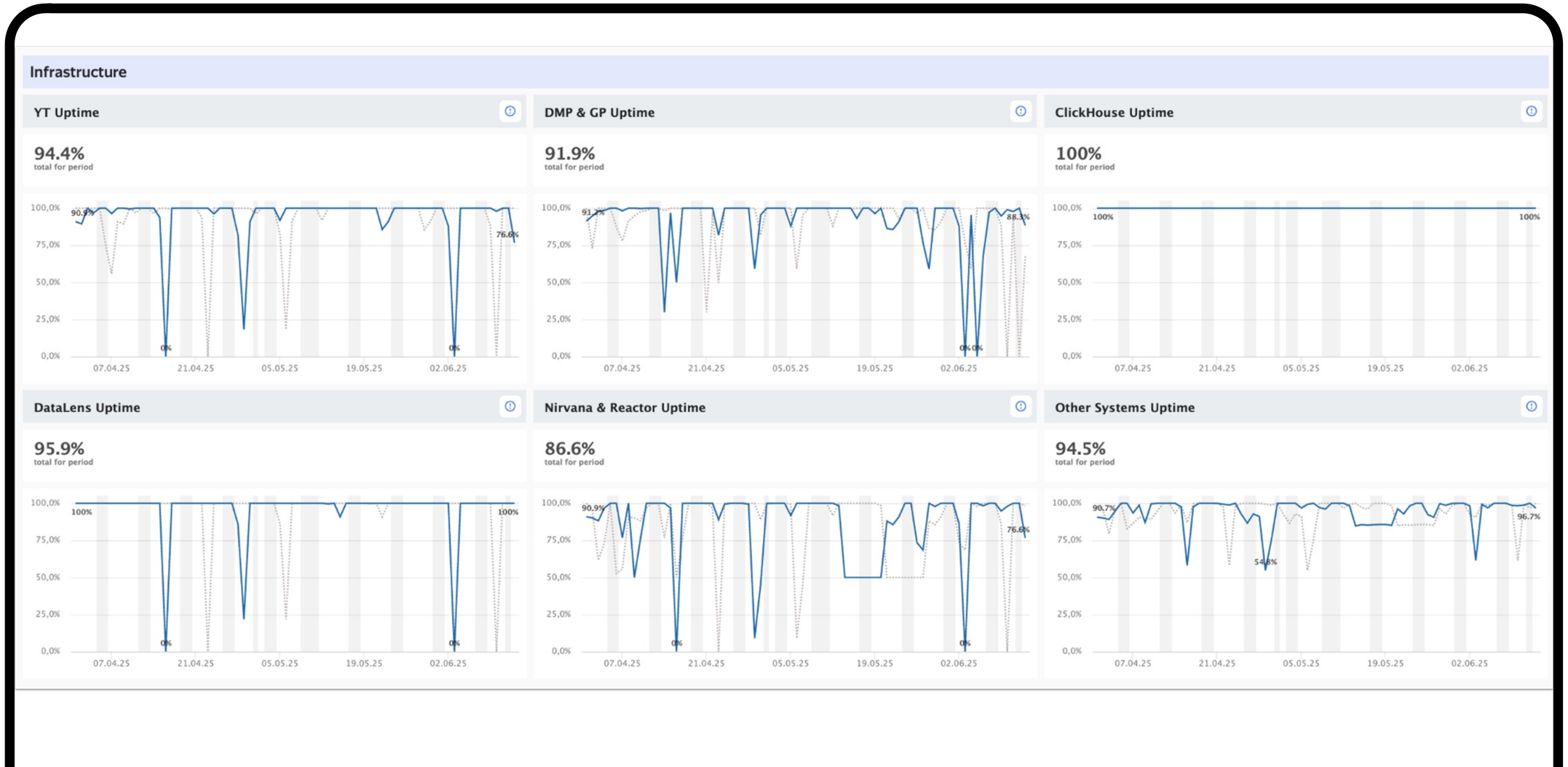
```
(ClickHouse + YT + DMP & GP + DataLens + Nirvana & Reactor + Other Services) / 6
```

Other Services:

- `Arc`
- `Arcadia CI`
- `Arcanum`
- `Atlas`
- `Codenv`
- `DC outage`
- `Juggler`
- `Logbroker`
- `MDB Core`
- `RTC`
- `Sandbox`
- `Transfer Manager`
- `YQL`



и делаем это в разрезе каждого сервиса, используемого в платформе данных Доставки



как создать для своей платформы метрику infrastructure Index

01 Определи ключевые архитектурные компоненты своей системы

02 Подключи метрики аптайма/health-check для каждого компонента
(Prometheis exporters, встроенные метрики из логов систем)

03 Сохраняй значения в витрину infrastructure_uptime
(Service, day, uptime_pct)

04 Считай индекс как средний uptime по всем сервисам за день
(HDFS + GP + Kafka + Spark + Airflow + Metabase)

05 Усредняй за период -> Infrastructure Index

06 В дашборде (Metabase/Superset) строй график тренда
и выделяй проблемные сервисы с командой

02

метрики стабильности платформы данных (SLA)

следим за выполнением гарантий обновления через Stability Index

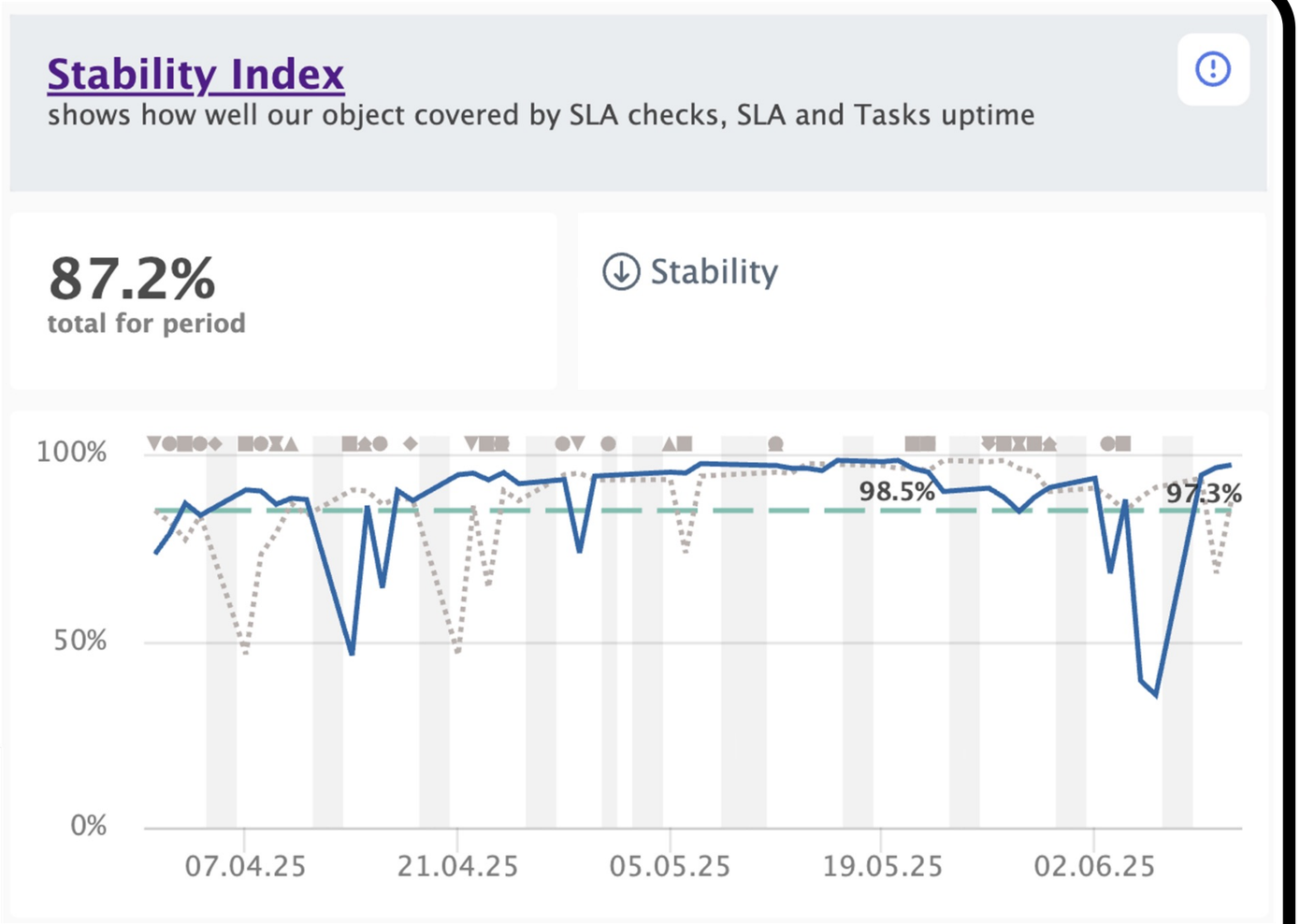
Stability Index

This metric shows the timeliness and success rate of object updates. It represents how well our objects are covered by SLA checks, whether our objects are updated within the established SLA, and the success rate of our data processing and update pipelines.

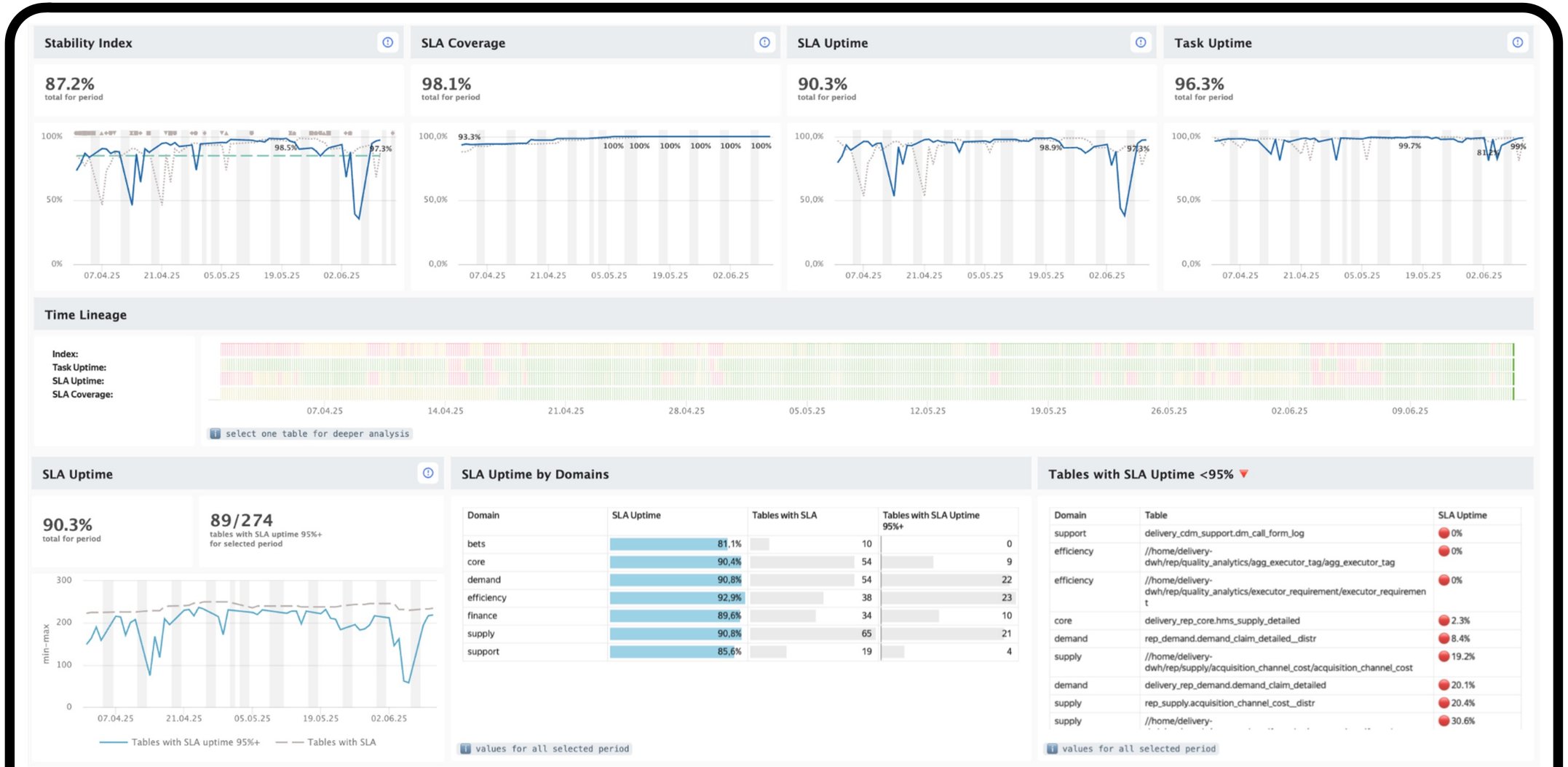
Index calculated as AVG stability by each table-day.

> **Formula:**
> `Share of time when SLA and Tasks was OK * has SLA flag`

**Index calculated for CDM,REP.*
Selector-like tables not included.



ключевой фокус в стабильности платформы — это прогнозируемое время обновления данных (SLA)



как создать для своей платформы метрику Stability Index

01 Определи SLA-контракты с бизнесом и аналитикой

02 Собери данные SLA и аптайма витрин из оркестраторов (Airflow + DBT, Cron, DMP)

03 Для таблиц с has_sla = true фиксируй периоды, когда обновления выполнялись в срок и успешно

04 Считай долю времени, когда SLA и задачи были OK (Stability = share_of_time(SLA OK & Tasks OK)

05 SI = AVG(Stability) по каждой таблице-дню -> Stability Index

06 В BI отобрази тренд индекса и список таблиц с отклонением от SLA

03

метрики качества данных

следим за качеством данных в нашей платформе через Quality Index

Quality Index

This metric shows the overall state of data quality. It represents how well our objects are covered by DQ checks and how successfully our objects pass these checks.

Index calculated as AVG value by each table-day.

> **Formula per 1 table:**
> `SUM (DQ Existing Measure Weight * DQ Check OK)`

> **For example:** table_1 has 2 DQ checks
> 1 check weight = 0.2, check ok =95%;
> 2 check weight = 0.05, check ok =80%
> **Index** = (0.2 * 0.95 + 0.05 * 0.8) = 23%
> ** more about check weight read in DQ Coverage description.*

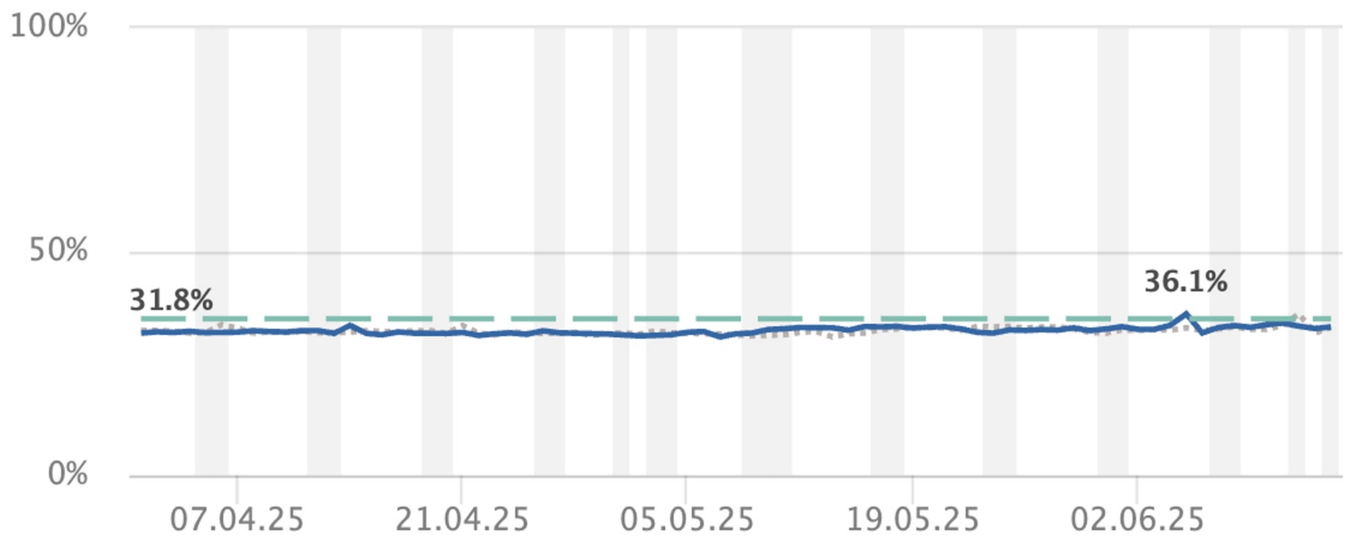
** Index calculated for CDM,REP.*
Selector-like tables not included.
Replicated tables not included

Quality Index

shows how well our object covered by DQ checks and how well this checks executed

32.4%
total for period

Quality



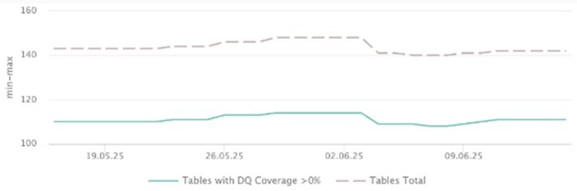
с помощью метрик качества данных мы прозрачно управляем задачами на инженеров в доменах

DQ Coverage

DQ Coverage

15 June 25
36.8%
last actual value
▲0.4 pp WoW

15 June 25
111/142
tables with DQ Coverage more than 0%



Coverage by Domains

Date	Domain	DQ Coverage	Tables Total	Tables with DQ Coverage >0%
15 June 25	bets	52,0%	5	5
15 June 25	core	30,8%	27	17
15 June 25	demand	42,0%	30	28
15 June 25	efficiency	23,5%	30	16
15 June 25	finance	48,8%	28	28
15 June 25	supply	35,8%	21	16
15 June 25	support	50,0%	1	1

values for last selected period

Tables without DQ ❌

Domain	Table	DQ Coverage
core	delivery_cdm_delivery.fct_claim_point	0%
core	delivery_rep_bi.agg_business_review	0%
core	delivery_rep_bi.agg_cancel_quality	0%
core	delivery_rep_bi.agg_marketplace_delivery	0%
core	delivery_rep_bi.dm_business_overview	0%
core	delivery_rep_bi.dm_finance_overview	0%
core	delivery_rep_metric_store.repeat_order_b2b_additional	0%
core	delivery_rep_metric_store.repeat_order_c2c	0%
core	rep_core.hms_supply_detailed__distr	0%

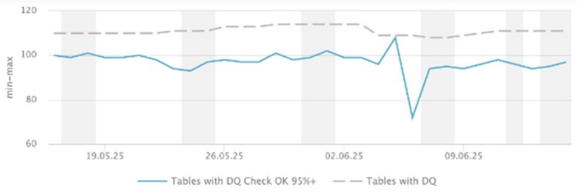
current state, not affected by date filter

DQ Check OK

DQ Check OK

96.1%
total for period

94/117
tables with DQ check ok 95%+ for selected period



DQ Check OK by Domains

Domain	DQ Check OK	Tables with DQ	Tables with DQ Check OK 95%+
bets	100.0%	5	5
core	92.1%	17	11
demand	96.3%	28	21
efficiency	96.1%	16	13
finance	98.8%	28	26
supply	93.7%	22	17
support	99.6%	1	1

values for all selected period

Tables with DQ Check OK <95% 🔻

Domain	Table	DQ Check OK
demand	delivery_rep_demand.demand_claim_detailed	50%
core	delivery_rep_cross_service_analytics.agg_delivery_metric_day	51.6%
supply	delivery_rep_supply.rep_supply_metric_bin	53.2%
supply	delivery_rep_supply.acquisition_channel_cost	66.7%
core	delivery_rep_bi.dm_delivery_overview	73%
demand	delivery_cdm_demand.dm_client_claim_daily	74.9%
efficiency	//home/delivery-dwh/cdm/dispatch/fct_route	75%
core	delivery_rep_bi.dm_finance_detailed	75.4%
efficiency	//home/delivery-dwh/cdm/dispatch/fct_edge	75.8%

values for all selected period

как создать для своей платформы метрику Quality Index

01 Установи Great Expectations и подключи свой DWH/S3 как datasource

02 Создай expectation suites для таблиц, укажи вес проверки в **meta.dq_weight**

03 Настрой checkpoint для запуска проверок

04 Сохраняй результаты в SQL-таблицу **dq_results** (check, вес, success_pct)

05 Считай индекс: QI = SUM(weight × success_pct) по таблице-дню, потом среднее

06 Собери дашборд: Quality Index, Coverage, Check OK, проблемные таблицы

07 Следи в динамике за этой метрикой и обсуждай
с командой в разрезе доменов на регулярной основе

04

метрики скорости отчётов

следим за скоростью отчетов платформы данных в DataLens через метрику

Dashboard Performance Index

Dashboard Performance Index

This metric shows the share of fast and successful events during BI production dashboard usage by users.

Formula:

`Fast and Successful Events/ Total Events`

*Event - one chart load, time for query execution in database

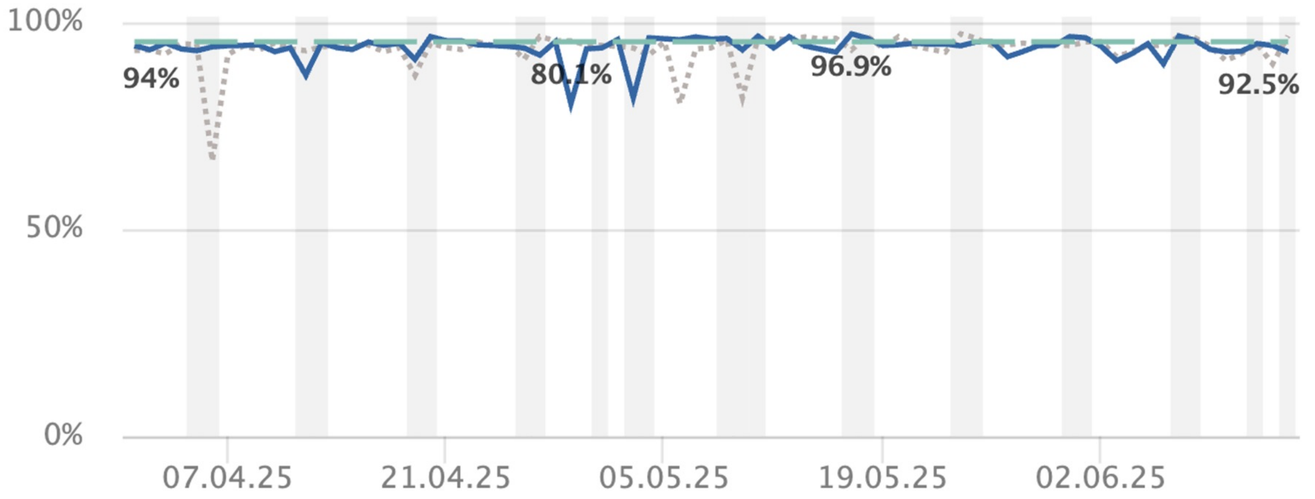
Dashboard Performance Index

shows how stable and fast BI dashboards work

93.6%

total for period

Dash Performance



как создать для своей платформы метрику Dashboard Perfomance Index

01 Настрой в своей BI системе логирование запросов

02 Собери витрину событий: **event_id, dashboard_id, chart_id, start_ts, end_ts, success_flag, type**

03 Задай пороги по типам: селекторы ≤0.5–3s, чарты ≤1–5s (определи с командой свои)

04 Рассчитай метрики по дню×дашборду: **total_events, fast_events, success_events**
(Fast% = fast_events/total_events, Success% = success_events/total_events)

05 Считай индекс: DPI = AVG (Fast% , Success%)

06 Сохраняй в **dashboard_perf_index** (dashboard_id, day, fast_pct, success_pct, dpi)

07 В BI собери дашборд: общий DPI, тренд Fast/Success, список проблемных дашбордов (низкий Fast%)

08 Следи в динамике за этой метрикой и обсуждай
с командой в разрезе доменов на регулярной основе

05

метрики скорости аналитических запросов

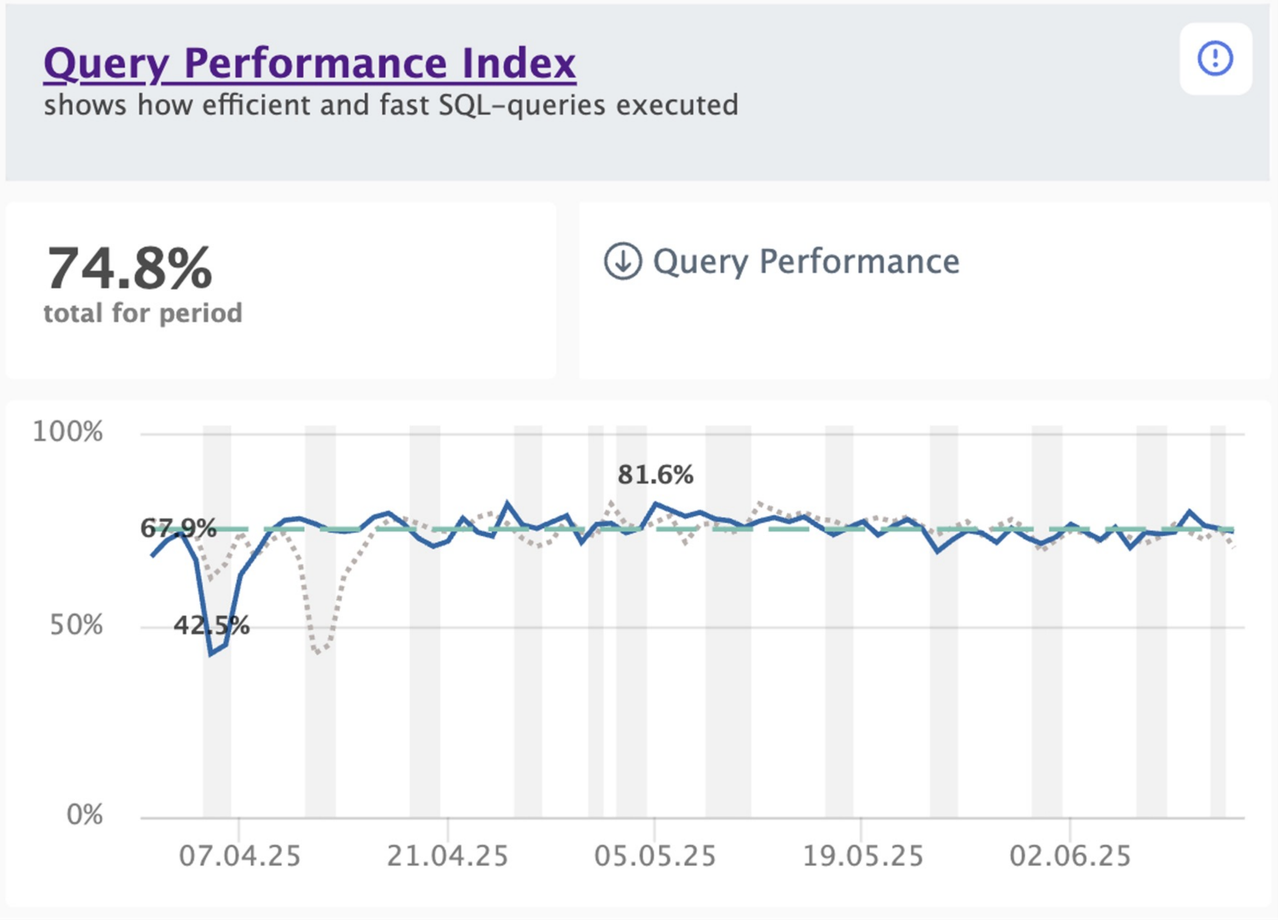
следим за скоростью аналитических запросов в разрезе доменов через метрику

Query Performance Index

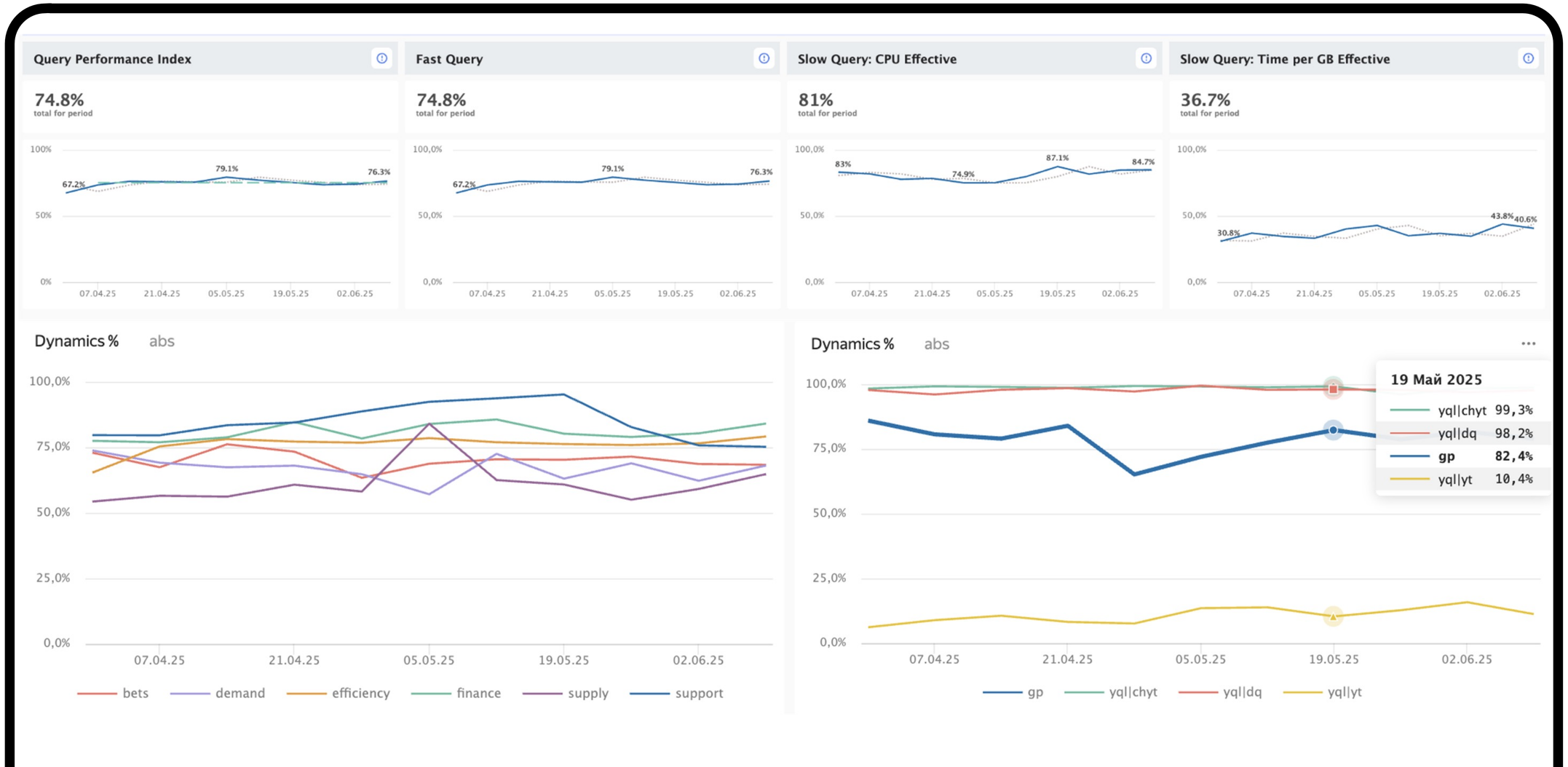
Query Performance Index

This index shows how fast and effective analytical scripts executed during ad-hocs or regular runs.

> **Formula:**
> `Fast and Effective Queries / Total Queries`



с помощью метрик скорости запросов выстраиваем линии поддержки аналитиков в каждом домене



как создать для своей платформы метрику Query Perfomance Index

01 Собирай на основе логов все пользовательские запросы в разрезе доменов/команд

02 Добавь метрики: duration = end_ts – start_ts, success_flag

03 Задай правила Fast and Effective
(например: runtime ≤30s для ad-hoc, ≤5m для batch, статус=успех)

04 Считай показатели по домену×дню : TotalQueries, FastEffectiveQueries, SystemType

05 QPI = FastEffectiveQueries / TotalQueries

06 Собери дашборд: Query Perfomance Index, Check OK, проблемные таблицы

07 Следи в динамике за этой метрикой и обсуждай
с командой в разрезе доменов на регулярной основе

06

метрики использования продуктов платформы данных бизнесом и аналитикой

следим за использованием отчётов в разрезе пользователей и доменов через метрику

Dashboard Usage Index

Dashboard Usage Index

This index shows proportion of weighted views to production dashboards created by BI team vs. all other dashboards in 05 Delivery.

Formula:

`BI Dashboard Weighted Views / Total Weighted Views`

Weighted views - view * coef:

- CEO = 100
- CEO -1 = 50
- CEO -2= 10
- Leads = 3
- Worker = 1
- Support = 0.3
- Robots = 0

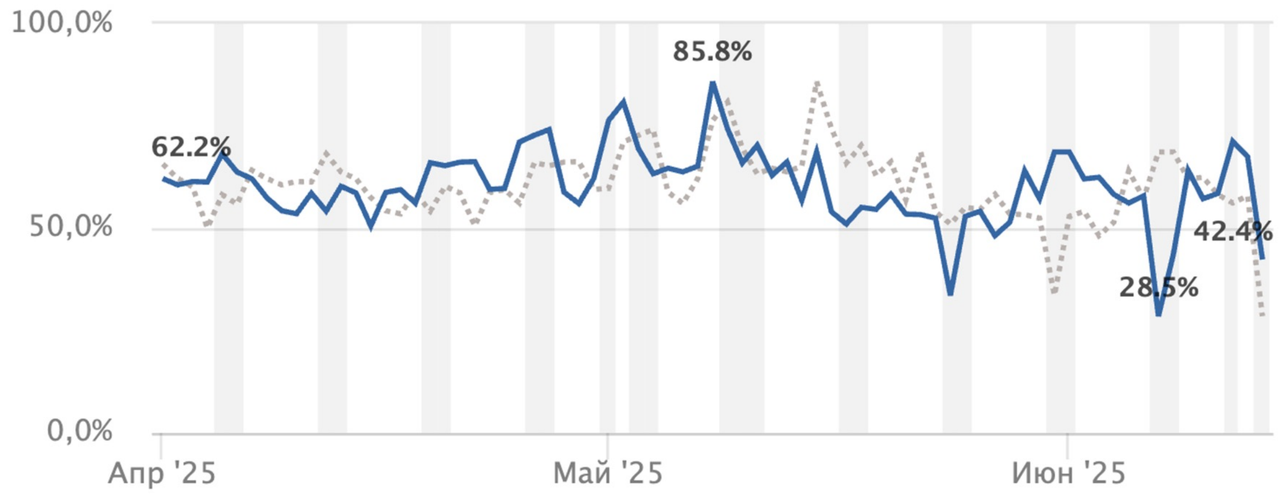
Dashboard Usage Index

shows the usage of data platform BI dashboards vs. all dashboards in 05Delivery

59.7%

total for period

Dashboard Usage



как создать для своей платформы метрику Dashboard Usage Index

01 Собери логи просмотров дашбордов (Metabase, Superset, PowerBI и так далее): user_id, dashboard_id, view_ts

02 Добавь веса ролей (пример: CEO=100, Head=50, Lead=10, Analyst=3, Worker=1, Support=0.3, Robots=0)

03 Посчитай Weighted Views: views × weight для каждой роли

04 Выбери те дашборды, которые достойны внимания и пользуются бизнесом

05 DUI = BI_Dashboard_Weighted_Views / Total_Weighted_Views

06 Сохраняй в витрину dashboard_usage_index (day, dui) и обновляй эти данные в t-1

07 Следи в динамике за этой метрикой и обсуждай с командой в разрезе доменов на регулярной основе

следим за использованием таблиц в разрезе пользователей и доменов через метрику

Table Usage Index

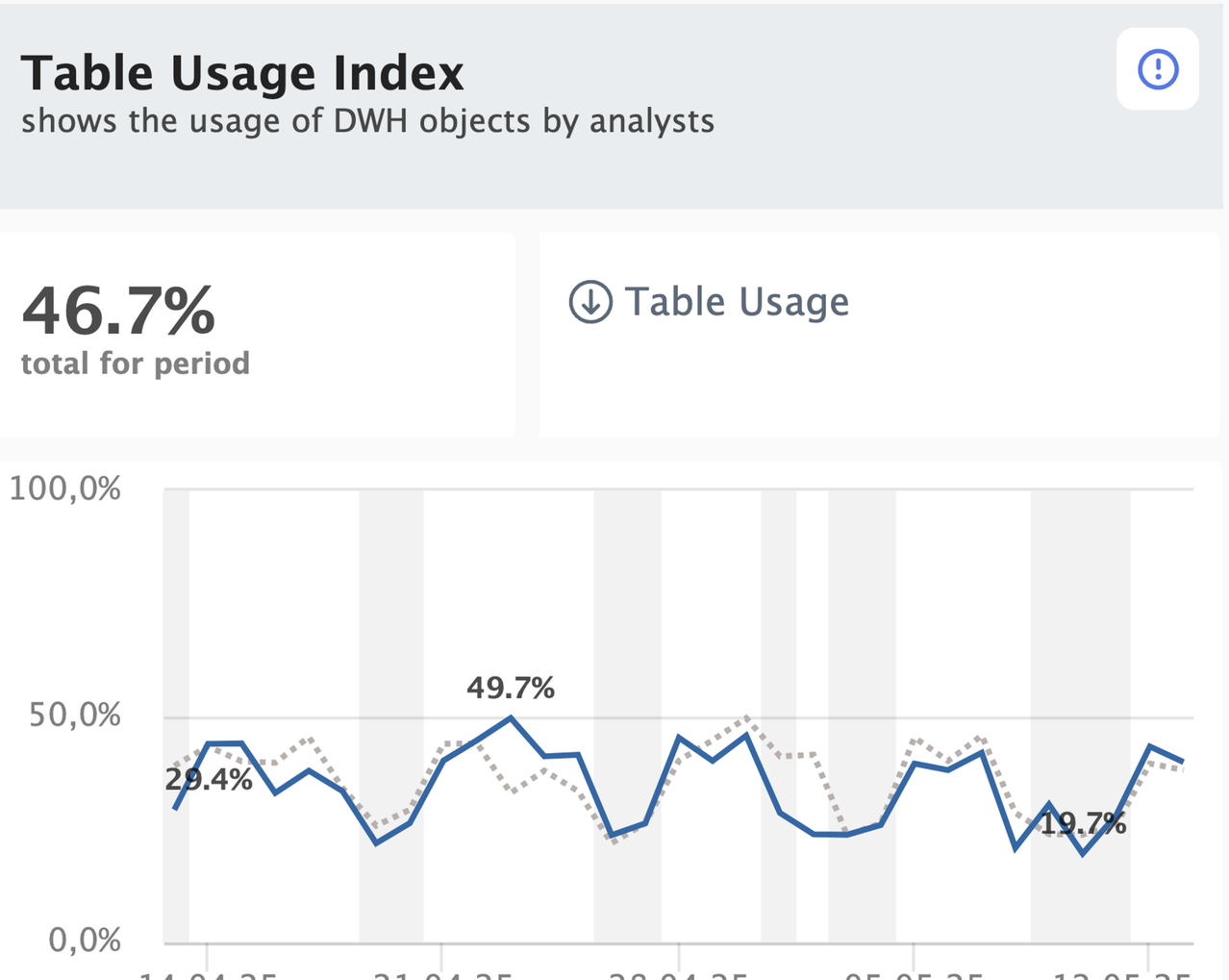
Table Usage Index

This metric shows the usage of DWH objects by analysts. It reflects the proportion of DWH tables used in analysts' work compared to all other objects, the share of active users, and a comparison of our DWH usage with DWHs from other business units.

Index calculated as weighted sum of 3 main measures.

Formula:

```
Delivery DWH Usage x0.6 + Active Users% x0.3 + Delivery vs Other DWH Usage x0.1
```



как создать для своей платформы метрику Table Usage Index

01 Собери события использования DWH: кто, когда и какую таблицу использовал

02 Посчитай долю запросов к вашему DWH /
запросы ко всем аналитическим данным в компании

03 Посчитай долю активных пользователей, определяй новичков/старичков и сравнивай их

04 Сравнивай сценарии использования, следи за тем, какие слои используют твои пользователи

05 Сохраняй в витрину dashboard_usage_index (day, dui) и обновляй эти данные в t-1

06 Следи в динамике за этой метрикой и обсуждай
с командой в разрезе доменов на регулярной основе

07

метрики эффективности использования ресурсов

следим за эффективностью использования ресурсов в разрезе доменов через метрику

Efficiency Index

Efficiency Index
This index shows approximate quota costs. (docs)

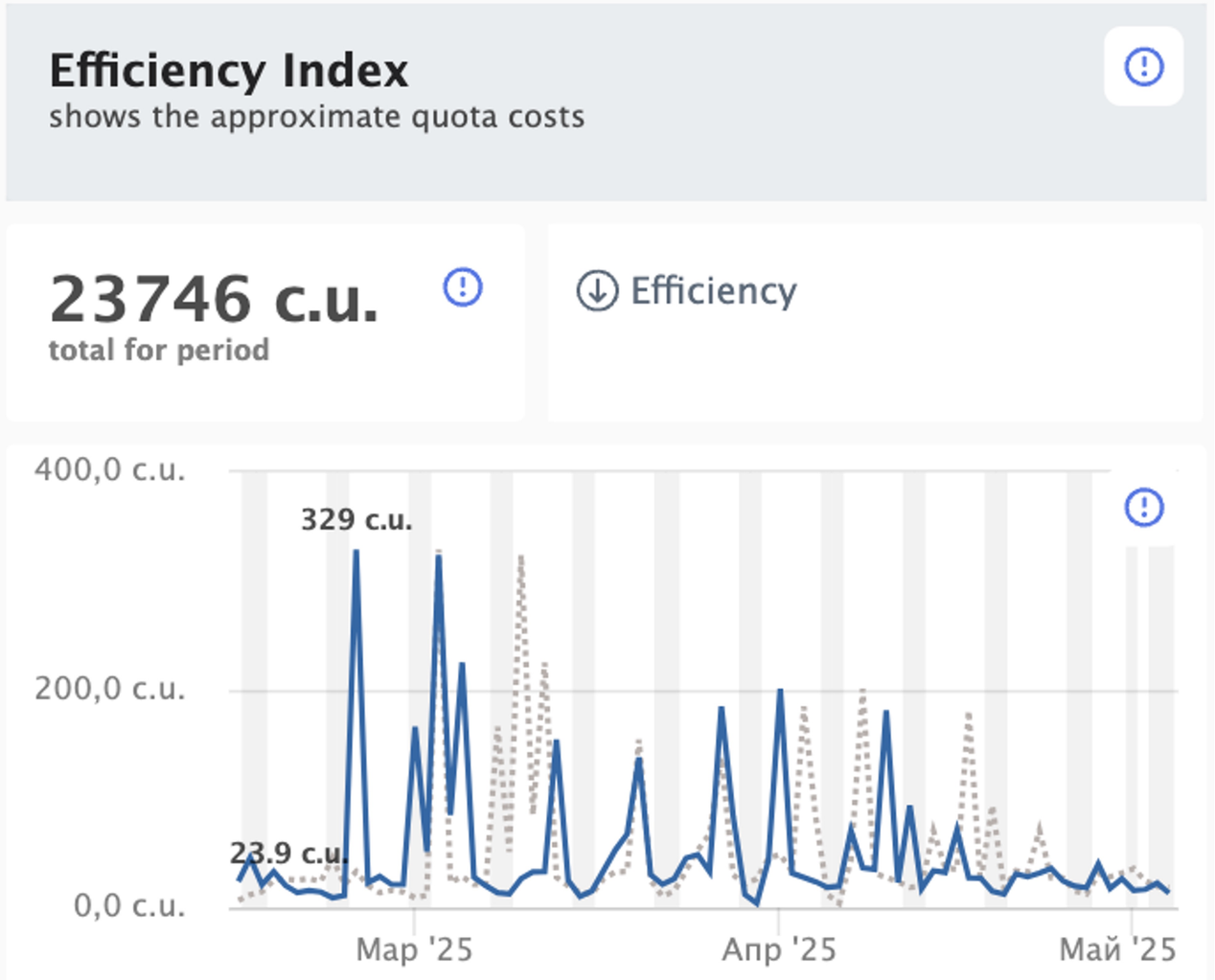
The cost of ownership is a guideline for estimating the approximate cost of a quota. It shows the monthly cost of the current quota volume at the moment.

The metric is displayed in conventional units (**c.u.**)

> **c.u.** - conventional units are a virtual currency for estimating and comparing the cost of ownership of resources.

ℹ️ *To make the metric more flexible, we recalculated it by days, where 1c.u. month = 1/30 c.u. per day*

> **Formula:**
> `SUM (Total c.u.)`



как создать для своей платформы метрику Efficiency Index

01 Определи квоты ресурсов (DWH, кластеры, хранилище, BI): CPU, RAM, Storage, Queries

02 Назначь «стоимость» квоты в условных единицах (c.u.) — внутренняя виртуальная валюта для сравнения

03 Собери фактическое потребление квот по дням

04 Пересчитай в c.u.: если 1 квота = 30 c.u./месяц → тогда ≈1 c.u./день

05 EfficiencyIndex = sum (Total \ c.u.)

06 Сохрани в витрину efficiency_index (day, total_cu) и обновляй эти данные в t-1

07 Следи в динамике за этой метрикой и обсуждай с командой в разрезе доменов на регулярной основе

08

метрики инженерной культуры аналитиков

следим за инженерной культурой в разрезе доменов через метрику

Engineering Culture Index

Nirvana Engineering Culture Index

This index shows how analysts fits riles when creating pipelines in Nirvana.
Read more about rules here (wiki)

Formula:

`AVG( Number of rules followed/ 7) per pipeline`

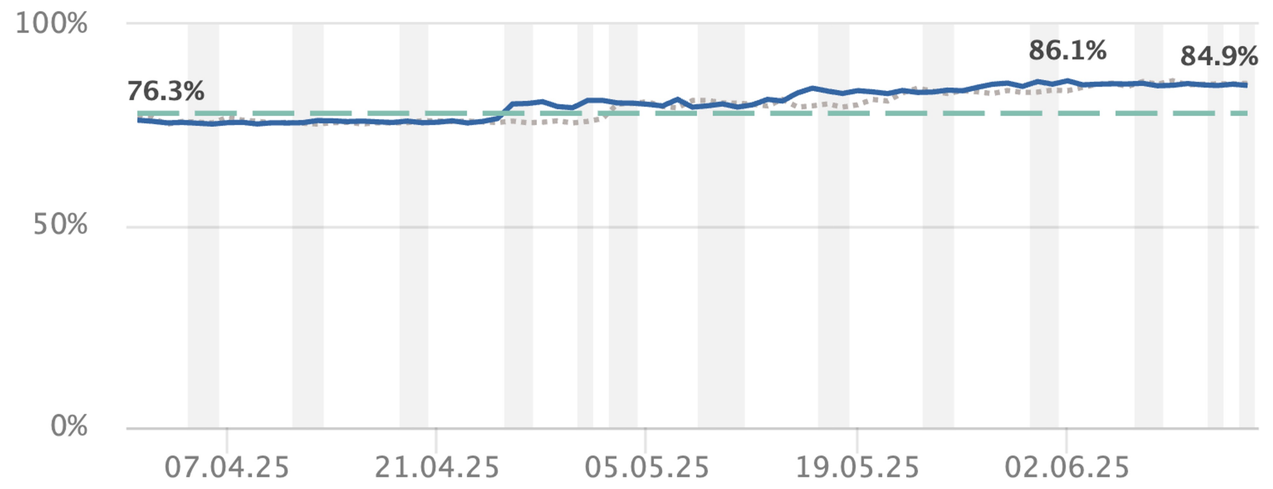
* 7- targeted number of rules

Engineering Culture Index

shows how well analysts follow rules when creating Nirvana pipelines

80.6%
total for period

Engineering Culture



как создать для своей платформы метрику Engineering Culture Index

01 Определи правила качества кода/пайплайнов
(например: тесты, логирование, мониторинг, документация, алерты, SLA, кодревью)

02 Логируй факт выполнения правил при создании ETL/ML/DA/BI пайплайнов

03 Для каждой задачи/пайплайна считай долю выполненных правил

04 ECI = AVG (pipeline\score)

05 Сохрани в витрину engineering_culture_index (domain, day, eci) и обновляй эти данные в t-1

06 Следи в динамике за этой метрикой и обсуждай с командой
в разрезе доменов на регулярной основе

07 Расширяй индекс — добавь использование инженерных инструментов
(Git PR, CI, тесты, мониторинг)

покрытие иерархии метрик

фреймворк необходимых действий по шагам

01

анализ бизнес-модели

02

декомпозиция
бизнес-модели
на доменные области

03

создание иерархии
метрик в каждом домене

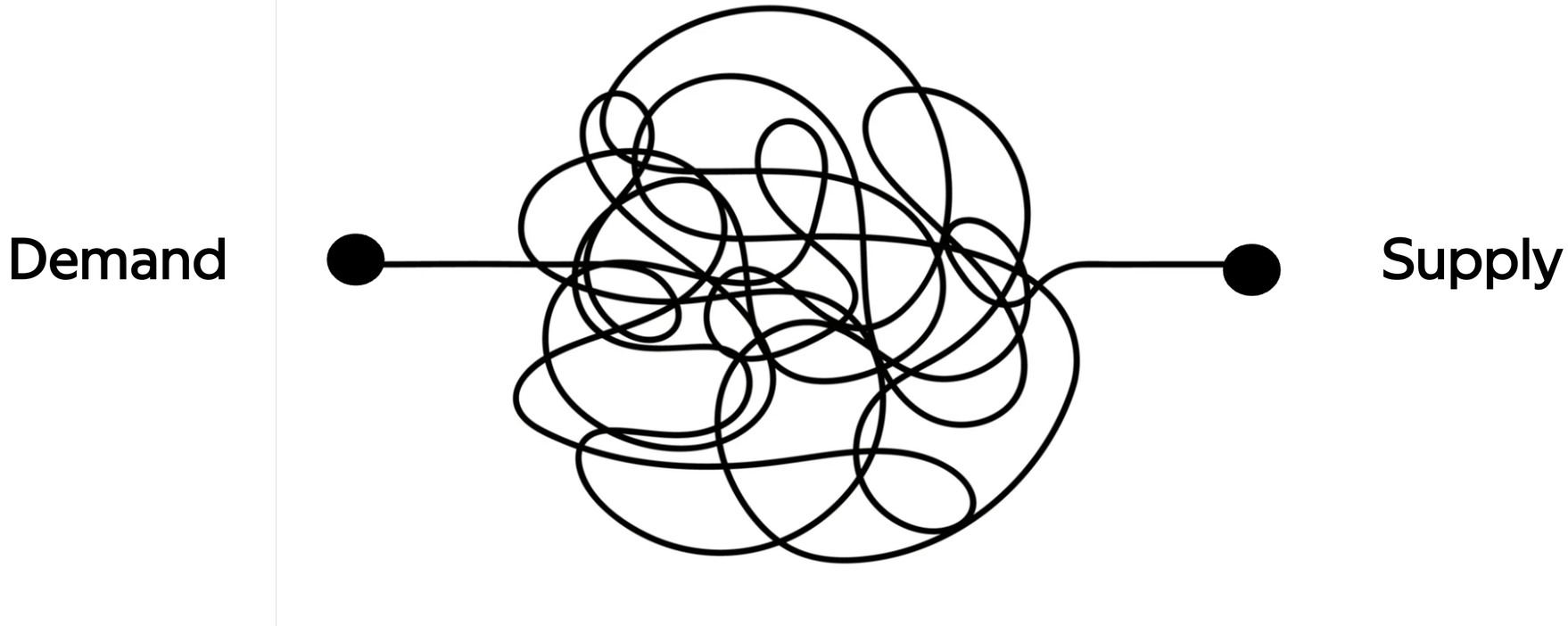
04

проектирование модели
данных и системы отчётности

05

развитие через % покрытия
иерархии метрик

анализ бизнес-модели



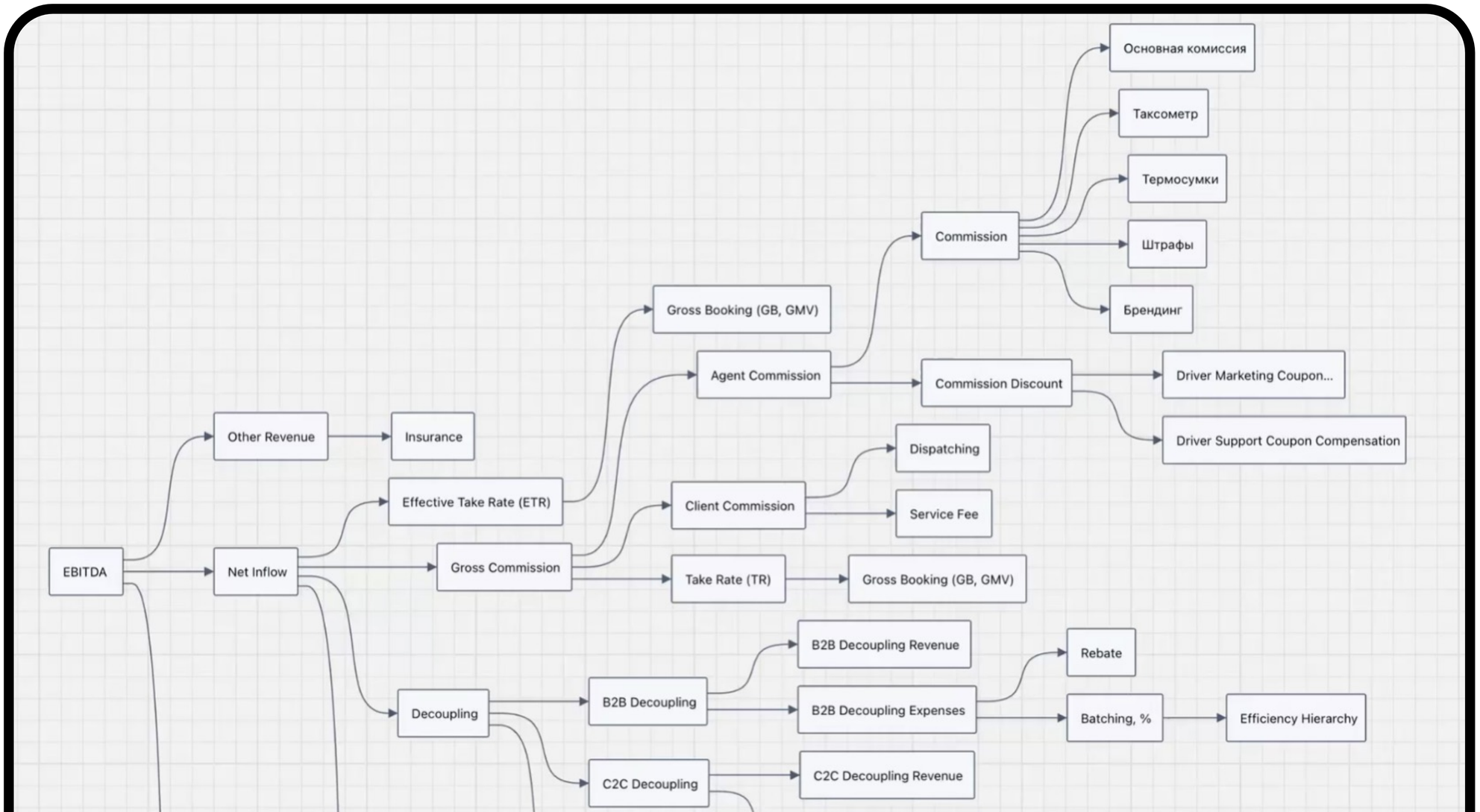
декомпозиция бизнес-модели на доменные области

Demand

Efficiency

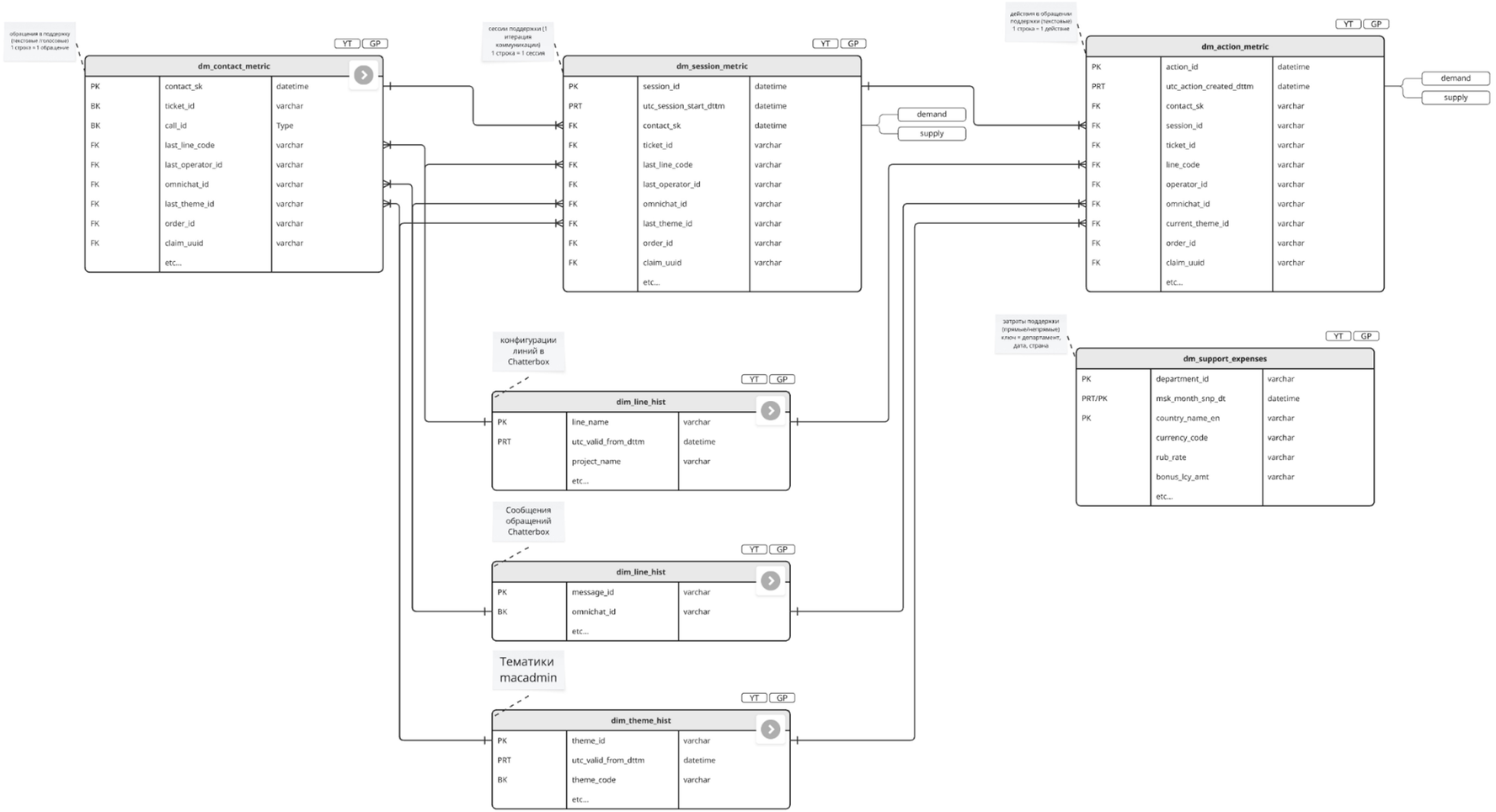
Supply

иерархии метрик в каждом домене бизнеса



проектирование gold-слоя

для аналитиков на основании метрик из иерархии бизнес-домена

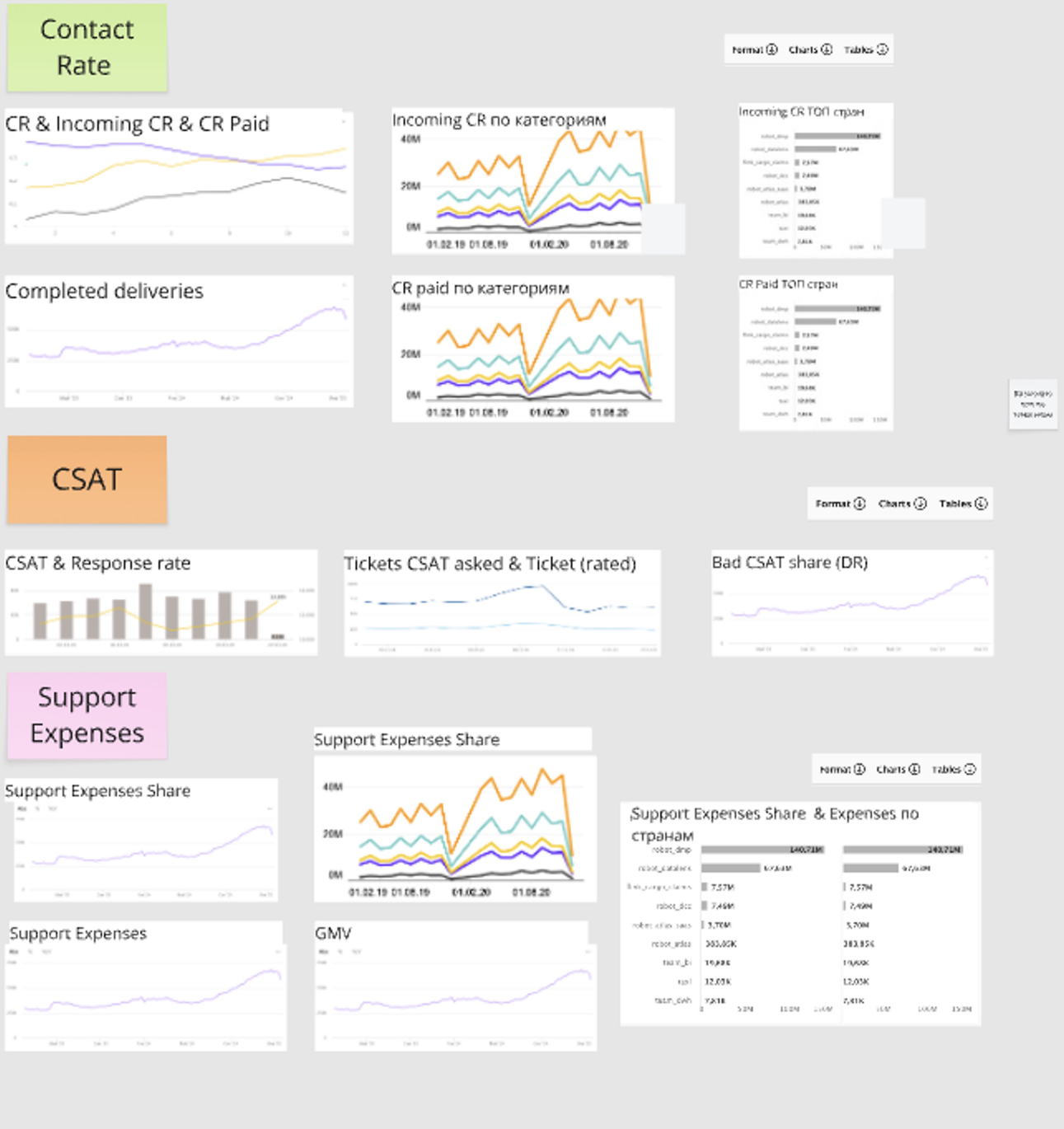


оценка качества модели аналитического слоя домена

через сложность и количество лишних соединений

проектирование системы отчётности домена на основании метрик из иерархии

проектирование системы отчётности домена на основании метрик из иерархии



ГОРОДСКИЕ СЕРВИСЫ

следим за развитием покрытия иерархии метрик в системе отчётности через

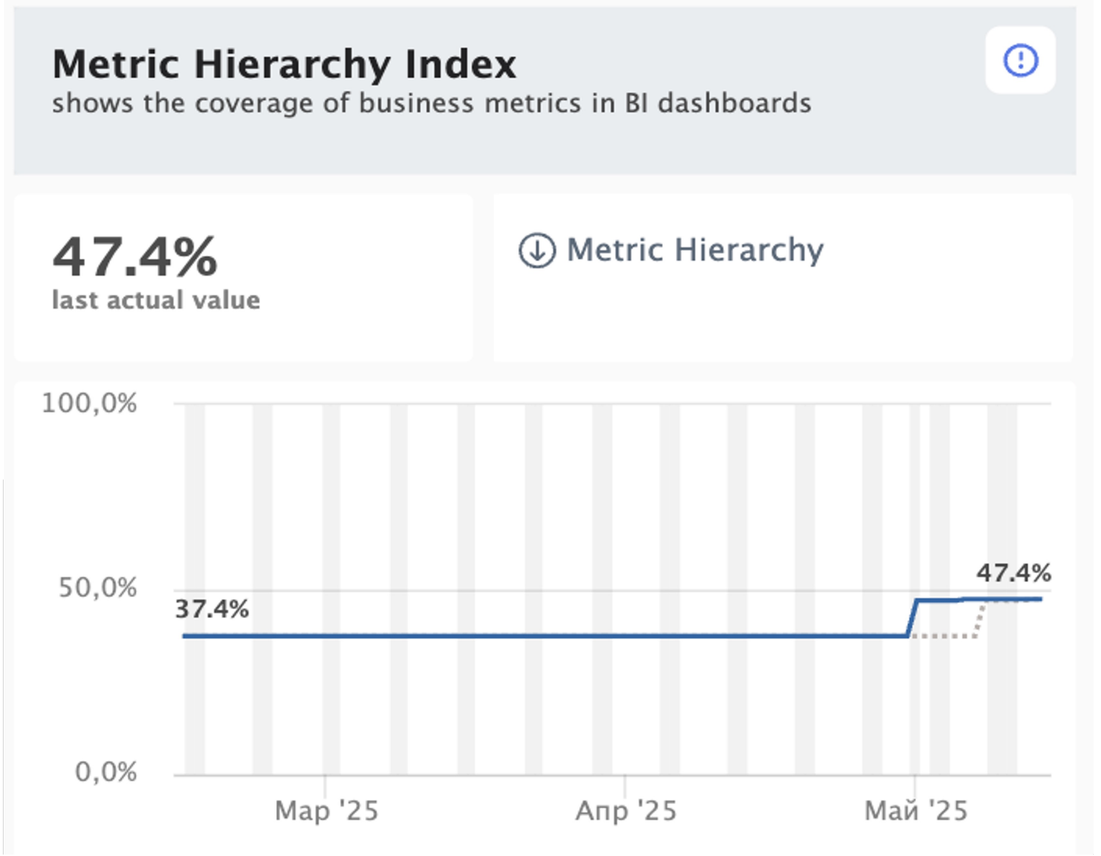
Metric Hierarchy Index

Metric Hierarchy Index

This index shows the coverage of business metrics from hierarchy that available for users in BI domain dashboards in DataLens

Formula:

```
Metrics available in DataLens /
Total metrics in hierarchy
```

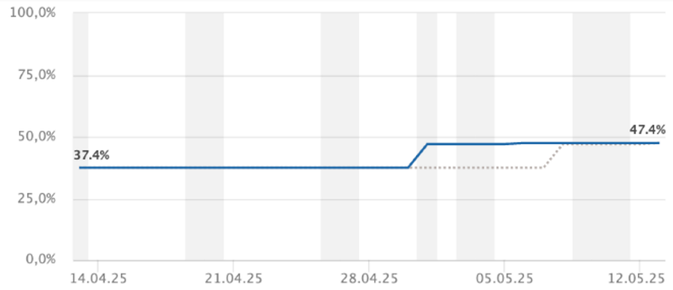


и делаем это
в разрезе каждого домена

Main Metrics

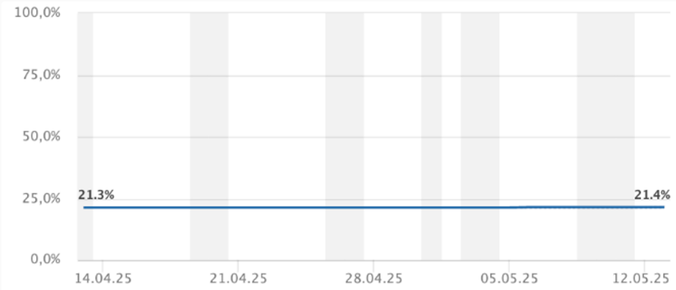
Metrics Available in BI Dashboards

47.4%
last actual value



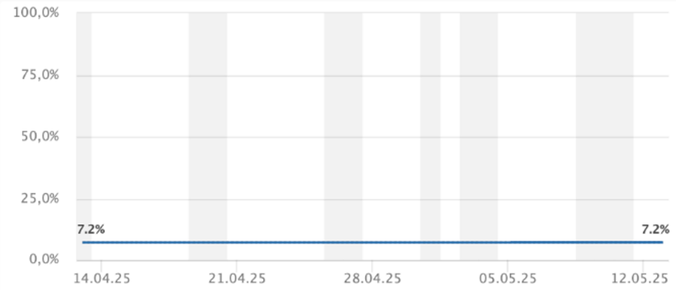
Metrics with Documentation

21.4%
last actual value



Metrics with Anomalies Detector

7.2%
last actual value



Metrics by Domains

Period	Hierarchy-> Metric	Metrics Total	Metrics available in DL	Metrics available in DL%	Metrics with documentation	Metrics with documentation%	Metrics with anomalies detector	Metrics with anomalies detector%	Metrics with description	Metrics with description%
13.05.2025	+ finance	77	59	77%	58	75%	13	17%	75	97%
13.05.2025	+ supply	59	39	66%	2	3%	1	2%	2	3%
13.05.2025	+ efficiency	184	99	54%	44	24%	21	11%	184	100%
13.05.2025	+ demand	103	33	32%	0	0%	0	0%	58	56%
13.05.2025	+ bets	62	0	0%	0	0%	0	0%	46	74%

как создать для своей платформы метрику Metric Hierarchy Index

01 Определи эталонную иерархию бизнес-метрик для каждого домена

02 Собери список метрик в BI (DataLens/Superset/Metabase), доступных пользователям в дашбордах

03 Сопоставь метрики BI с эталонной иерархией

04 Формула индекса: MHI = Metrics available in BI \ Total metrics in hierarchy

05 Агрегируй по своим доменам

06 Сохрани в витрину metric_hierarchy_index (domain, day, mhi)

07 Следи в динамике за этой метрикой и обсуждай с командой в разрезе доменов на регулярной основе

ВЫВОДЫ

ВЫВОДЫ

01

оцифровка компонентов позволяет создавать понятную для всех в бизнесе стратегию развития платформы данных

ВЫВОДЫ

02

для каждого из направлений платформы появляются оцифрованные цели и результаты

ВЫВОДЫ

03

инженеры понимают, на какие метрики и как именно они влияют через свои доработки и изменения в сценариях обработки данных

ВЫВОДЫ

04

системные аналитики понимают, как их объектная модель используется и могут делать регулярные выводы, как её улучшать

ВЫВОДЫ

05

bi-аналитики могут следить за развитием своей системы отчётности через покрытие иерархий метрик и метрик использования дашбордов

ВЫВОДЫ

06

всё это работает вместе как единая система с общими целями и создает правильную культуру внутри команд платформы данных

ВЫВОДЫ

07

в эту систему метрик заезжают платформы данных Такси, Еды, Лавки, Финтех, СИБ

как мы её прокачали, расскажу в следующих сериях

ГОРОДСКИЕ СЕРВИСЫ

Smart Data

готов ответить на вопросы

Влад Гоцуляк

Руководитель платформы данных в Яндекс Доставке