Москва | 2024

От ручного труда к автоматической генерации проверок качества данных

Безопасный транспорт

ЦОДД

ИЦ – ЧАСТЬ КОМАНДЫ ТРАНСПОРТНОГО КОМПЛЕКСА

Инновационный центр «Безопасный транспорт» создан **в 2017 году** для работы с Big Data и решения задач Транспортного комплекса Москвы

Безопасный транспорт

ЦОДД

Московский транспорт

ЦОДД

Московский метрополитен

Мосгортранс

Московский каршеринг

Такси

Московский паркинг

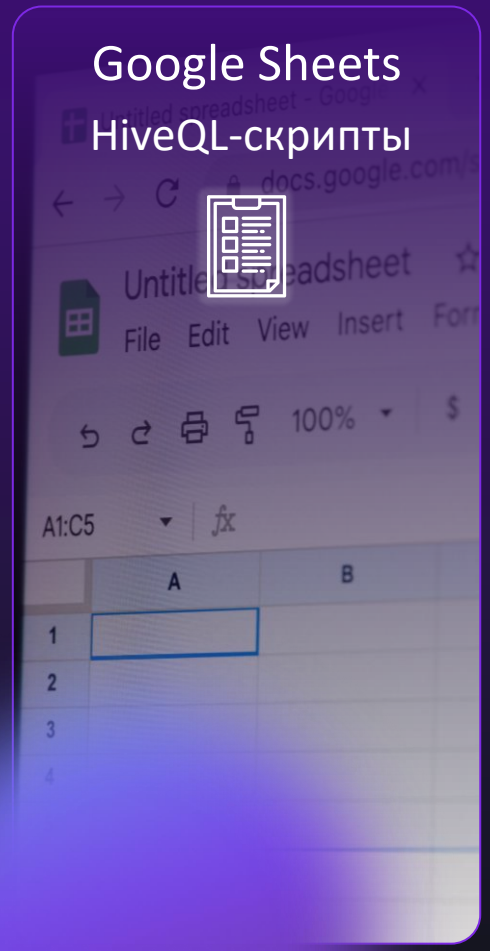
Велотранспорт

Ручная работа DQ

Аналитик

Google Sheets

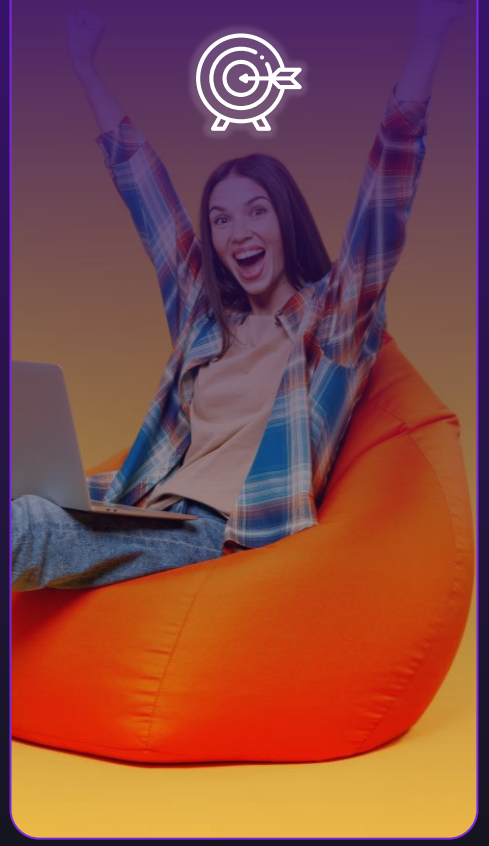
HiveQL-скрипты



DWH

HiveQL-скрипты

Результат

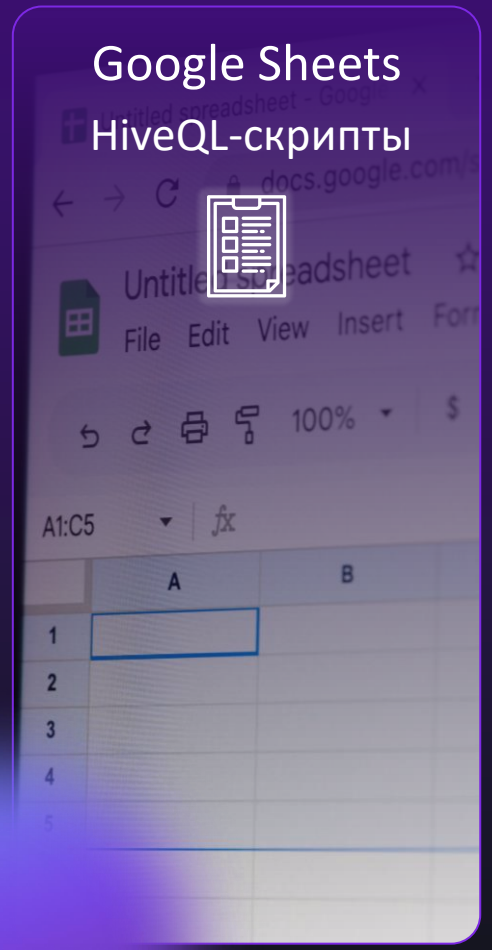


Автоматический запуск проверок DQ

Аналитик

Google Sheets

HiveQL-скрипты

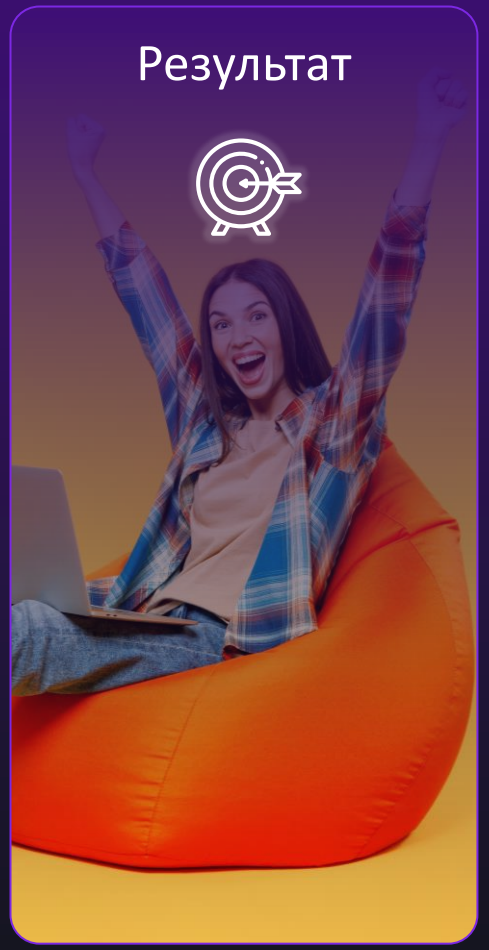


DQ Launcher

DWH

HiveQL-скрипты

Результат



Проблемы на поверхности

Проверки не для всех таблиц

Устаревание проверок при изменении таблиц

Огромный объём monkey job

Оповещения на почту, которую нельзя подключить к телефону

Логические ошибки в проверках

Наложение выполнения проверок на обновление данных

Задержки в получении результатов проверок

Отдельные сотрудники на мониторинг DQ

Постоянная доработка дашбордов

Безопасный транспорт ЦОДД

Увеличение источников и продуктов

ЦОДД

Московский Метрополитен

Мосгортранс

Такси

Московский каршеринг

Велотранспорт

И многие другие...

Данные из **45+ систем** подведомственных организаций, в т.ч. **5 основных поставщиков** передают данные в близком к realtime режиме

40+ собственных экспертно-созданных справочников

Hadoop

Vertica

Apache Spark

IBM Campaign

Apache Hive

DB2

Apache Kafka

ClickHouse

Airflow

Postgres

Bash

Python

Хранилище Больших данных ТК

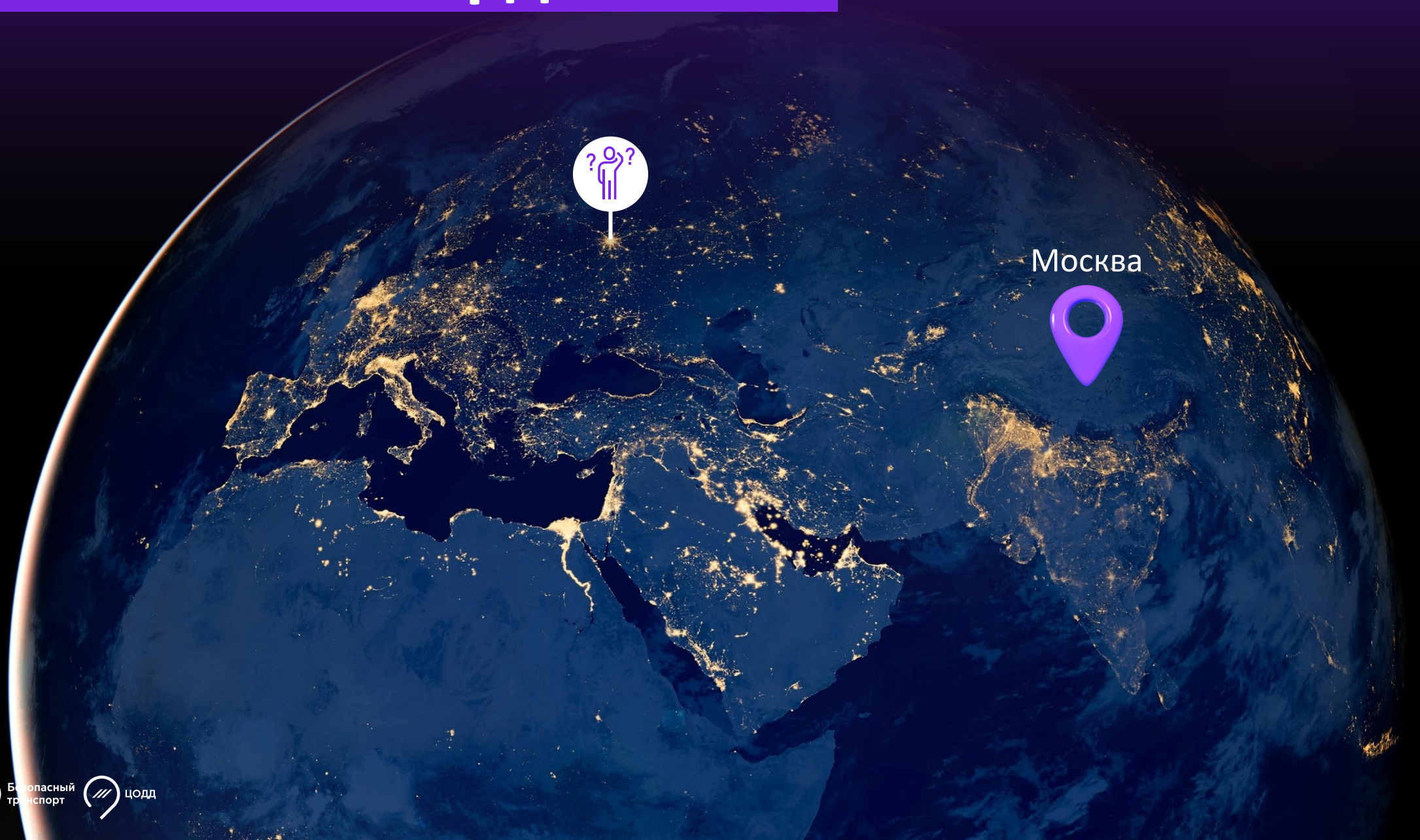
Обработка **50+Гб данных** каждый день

10+ моделей прогнозной аналитики

... и немного магии

Сервисы для горожан и аналитические продукты для руководства ЦОДД и ДТ

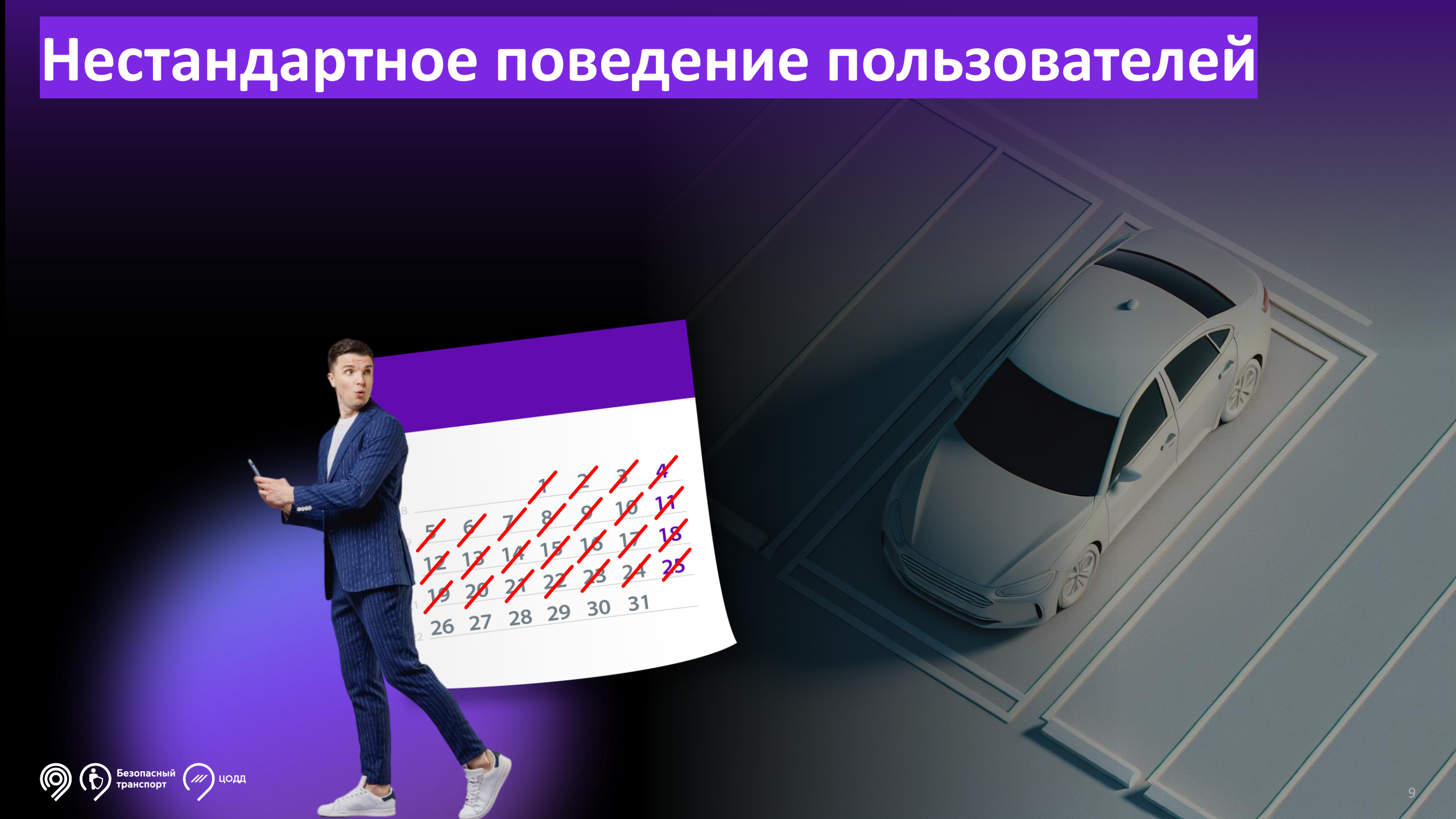
Кейсы с GPS-координатами



Нестандартное поведение транспортных средств

Безопасный транспорт ЦОДД

Нестандартное поведение пользователей



Безопасный транспорт ЦОДД

Data Lake

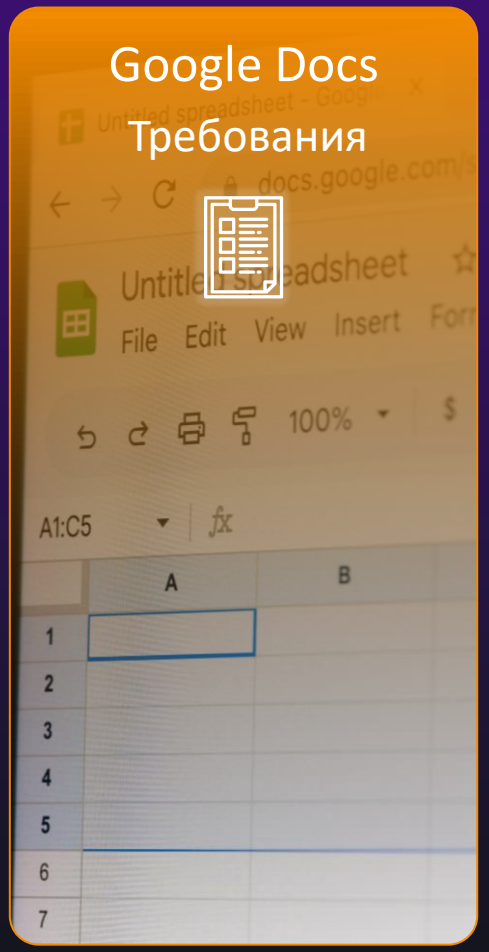
Спасение утопающих…

Автоматическая генерация проверок DQ

Аналитик

Google Docs

Требования



DQ metric Generator

HiveQL-скрипты

- наличие записей
- обязательные поля
- дубли
- выбросы

DQ Launcher

Результат

DWH

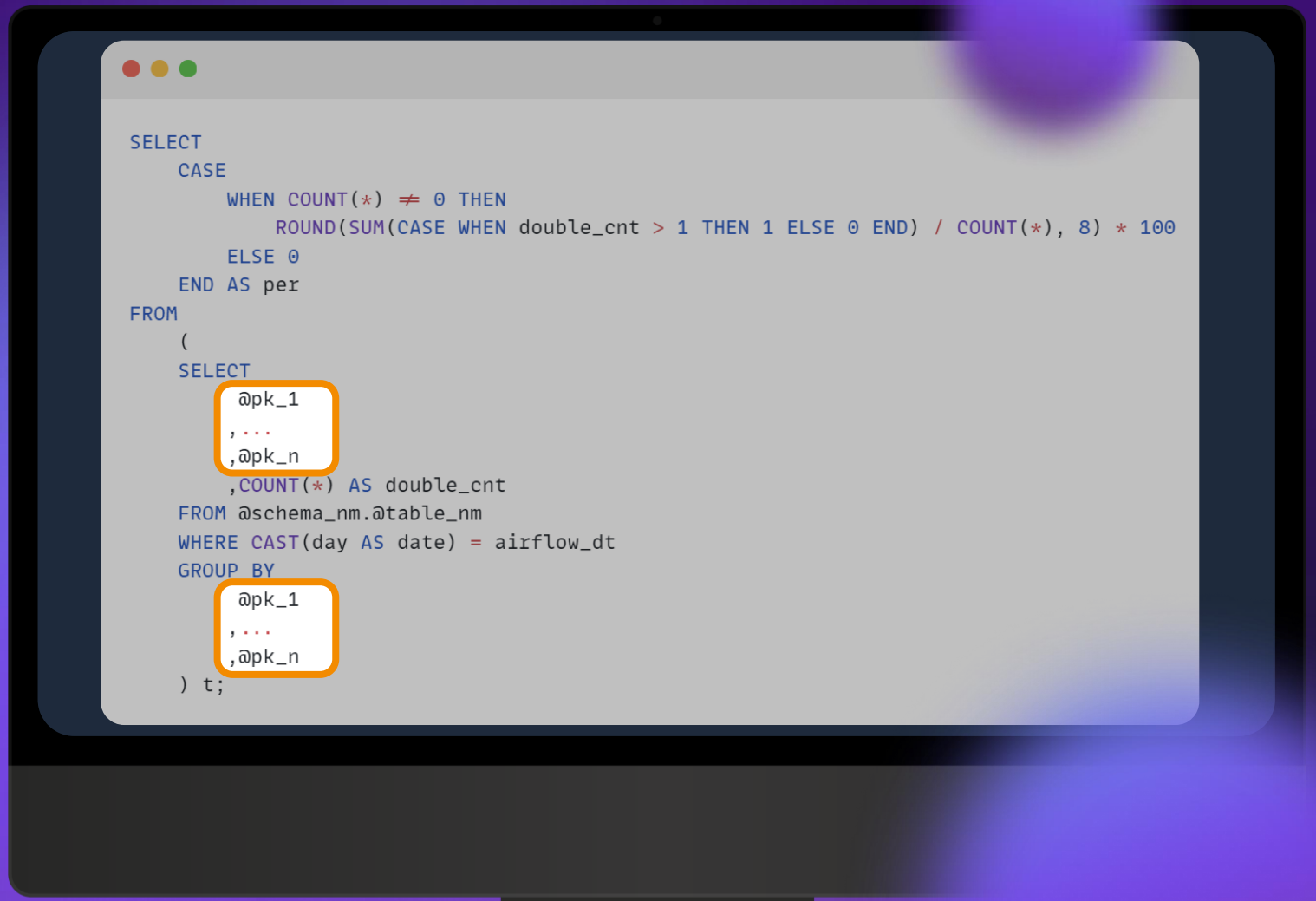
HiveQL-скрипты

Безопасный транспорт ЦОДД

Структурированный комментарий к атрибуту

```
CREATE TABLE `cifra_calc.s_parking_session_x_district_x_day_time`(
  `dt` date COMMENT 'PK | NotNull | Отчётная дата',
  `day_time` string COMMENT 'PK | NotNull | Время суток',
  `district_nm` string COMMENT 'PK | Район',
  `division_nm` string COMMENT 'PK | NotNull | Округ',
  `payment_way_nm` string COMMENT 'PK | NotNull | Способ оплаты',
  `parking_session_cnt` int COMMENT 'NotNull | Количество парковочных сессий',
  `parking_session_per` double COMMENT 'NotNull | Процент парковочных сессий, оплаченных конкретным способом оплаты',
  `parking_session_avg_dur` double COMMENT 'NotNull | Средняя длительность парковочной сессии, мин',
  `vehicle_cnt` int COMMENT 'NotNull | Количество уникальных ГРЗ',
  `period_from_dt` date COMMENT 'NotNull | Дата начала действия среза (техн.)',
  `period_to_dt` date COMMENT 'NotNull | Дата окончания действия среза (техн.)',
  `process_dttm` timestamp COMMENT 'NotNull | Дата и время добавления записи в таблицу (техн.)')
PARTITIONED BY (
  `day` string COMMENT 'NotNull | День - партиция (техн.)')
```

Универсальный шаблон для генерации проверок на дубли



```
SELECT
    CASE
        WHEN COUNT(*) =/= 0 THEN
            ROUND(SUM(CASE WHEN double_cnt > 1 THEN 1 ELSE 0 END) / COUNT(*), 8) * 100
        ELSE 0
    END AS per
FROM
    (
    SELECT
        @pk_1
        ,...
        ,@pk_n
        ,COUNT(*) AS double_cnt
    FROM @schema_nm.@table_nm
    WHERE CAST(day AS date) = airflow_dt
    GROUP BY
        @pk_1
        ,...
        ,@pk_n
    ) t;
```

Безопасный транспорт ЦОДД

Первые успехи

Проверки не для всех таблиц

Устаревание проверок при изменении таблиц

Оповещения на почту, которую нельзя подключить к телефону

Огромный объём monkey job

Логические ошибки в проверках

Наложение выполнения проверок на обновление данных

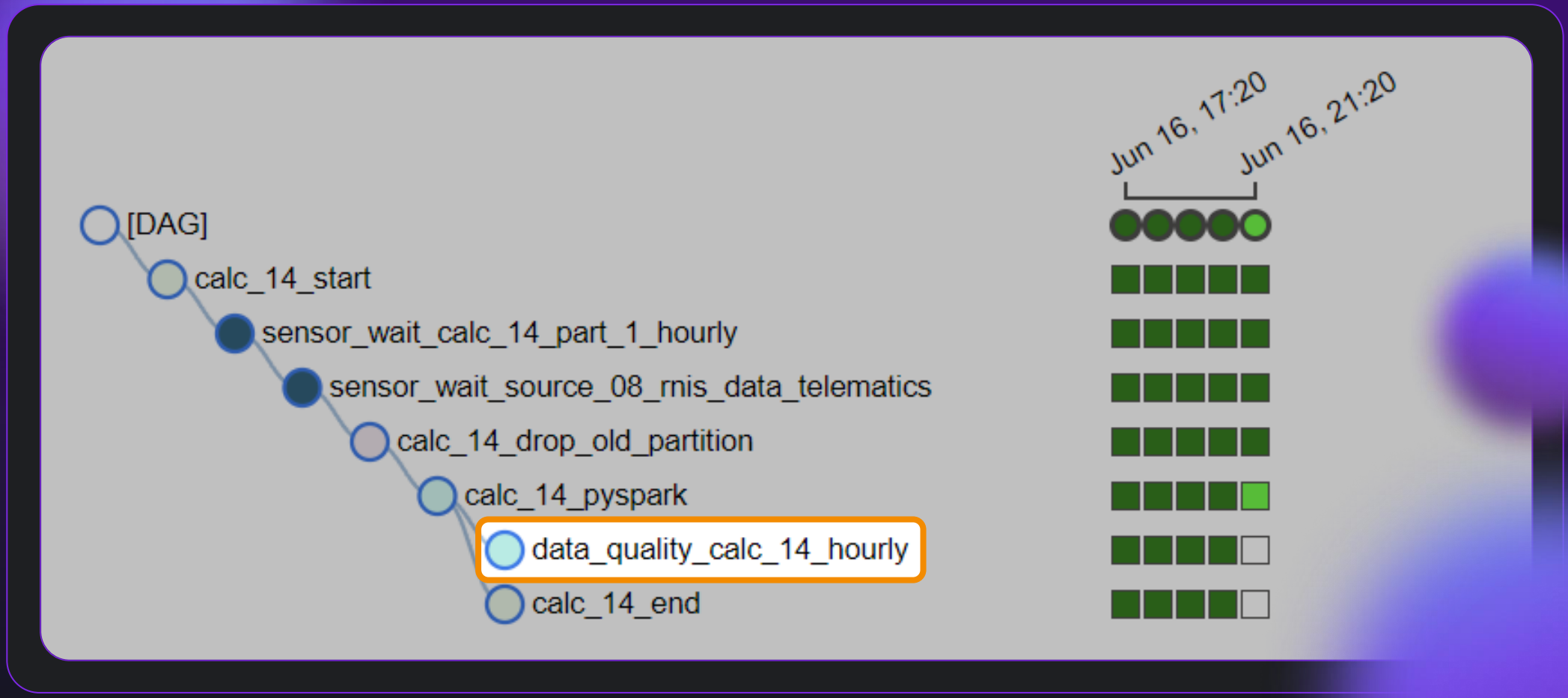
Задержки в получении результатов проверок

Отдельные сотрудники мониторинг DQ

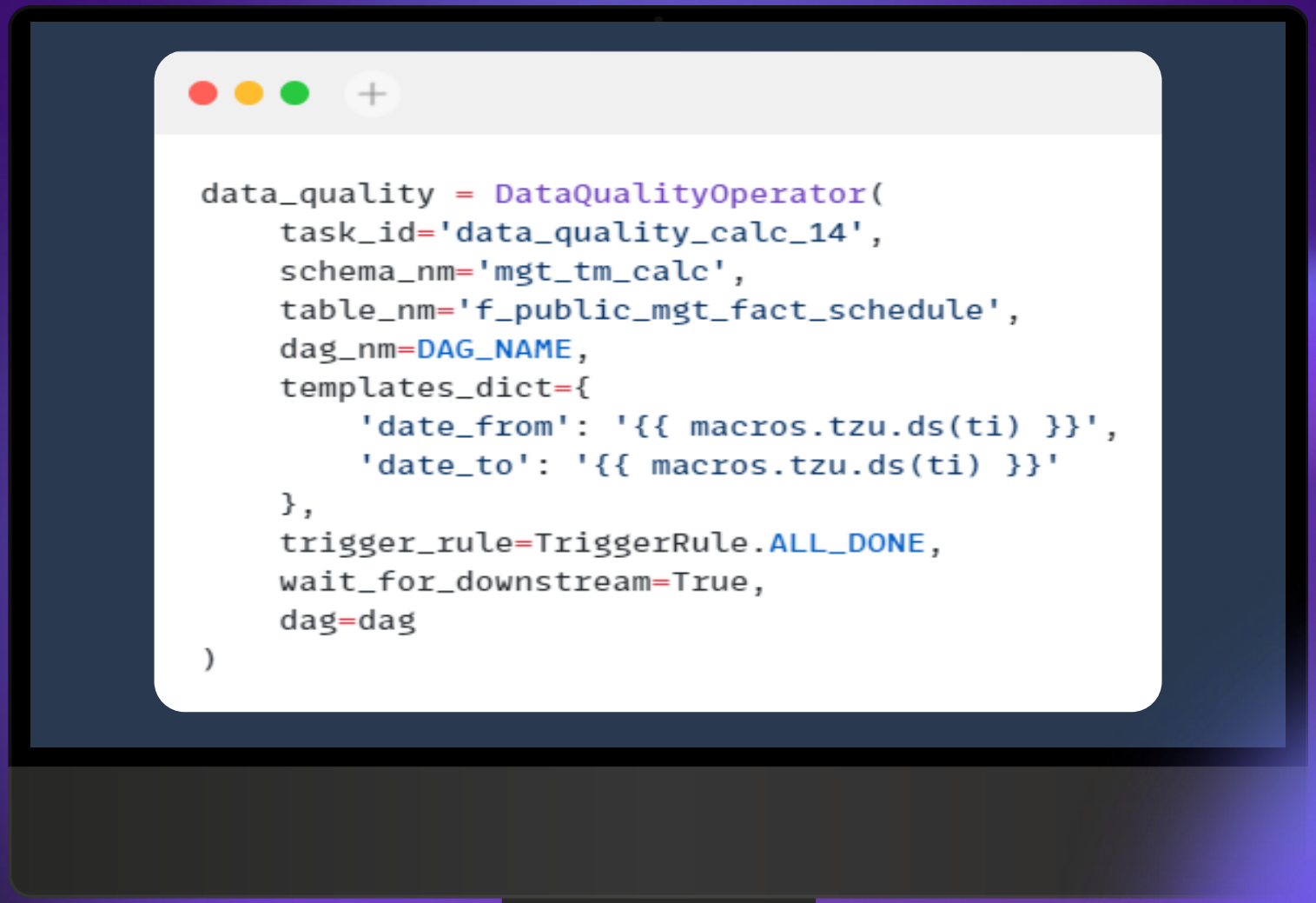
Постоянная доработка бордов

Безопасный транспорт ЦОДД

Решение вопроса задержки получения информации



Airflow Task



```
data_quality = DataQualityOperator(
    task_id='data_quality_calc_14',
    schema_nm='mgt_tm_calc',
    table_nm='f_public_mgt_fact_schedule',
    dag_nm=DAG_NAME,
    templates_dict={
        'date_from': '{{ macros.tzu.ds(ti) }}',
        'date_to': '{{ macros.tzu.ds(ti) }}'
    },
    trigger_rule=TriggerRule.ALL_DONE,
    wait_for_downstream=True,
    dag=dag
)
```

Очередные успехи

Проверки не для всех таблиц

Устаревание проверок при изменении таблиц

Оповещения на почту, которую нельзя подключить к телефону

Огромный объём monkey job

Логические ошибки в проверках

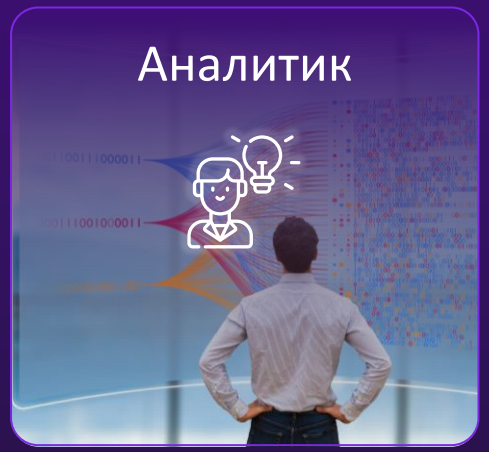
Наложение выполнения проверок на обновление данных

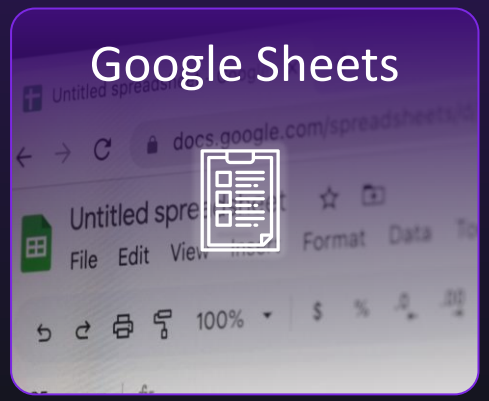
Задержки в получении результатов проверок

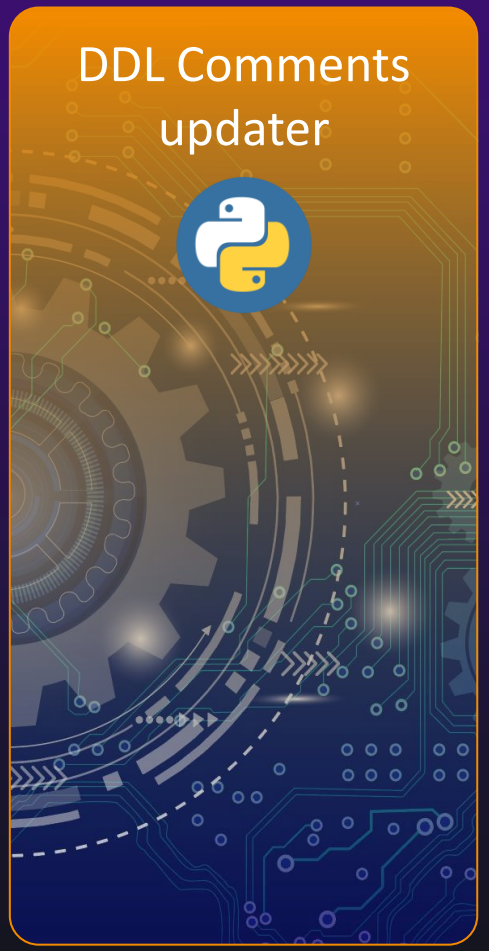
Отдельные сотрудники на мониторинг DQ

Постоянная доработка бордов

Автоматическое обновление DDL комментариев







Безопасный транспорт ЦОДД

Ведение свойств атрибутов в требованиях

Номер витрины	Код схемы	Код таблицы	Код атрибута	Описание атрибута	Тип данных в Hive	Обязательное	Вхождение в первичный ключ	Измерение Dimension	Показатель Measure
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	DT	Отчётная дата	date	+	+	+	
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	WORKING_DAY_FLG	Признак рабочего дня	int	+		+	
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	WEEK_NUM	Номер отчётной недели	int	+		+	
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	OPERATOR_NM	Перевозчик	string	+	+	+	
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	VEHICLE_CNT	Фактическое количество ТС (выпуск)	int	+			+
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	PLAN_VEHICLE_CNT	Плановое количество ТС (выпуск)	int	+			+
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	COMPLETED_VEHICLE_PERCENT	Процент выполнения выпуска	double	+			+
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	TRIP_CNT	Фактическое количество рейсов	int	+			+
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	PLAN_TRIP_CNT	Плановое количество рейсов	int	+			+
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	COMPLETED_TRIP_PERCENT	Процент выполнения рейсов	double	+			+
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	PROCESS_DTTM	Дата и время добавления записи в таблицу (техн.)	timestamp	+			
161	CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	DAY	День - партиция (техн.)	string	+			
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	DT	Отчётная дата	date	+	+	+	
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	WORKING_DAY_FLG	Признак рабочего дня	int	+		+	
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	WEEK_NUM	Номер отчётной недели	int	+		+	
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	ROUTE_ID	Идентификатор маршрута	bigint	+	+	+	
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	ROUTE_NM	Наименование маршрута	string	+		+	
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	TRIP_CNT	Фактическое количество рейсов	int	+			+
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	PLAN_TRIP_CNT	Плановое количество рейсов	int	+			+
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	COMPLETED_TRIP_PERCENT	Процент выполнения рейсов	double	+			+
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	PROCESS_DTTM	Дата и время добавления записи в таблицу (техн.)	timestamp	+			
170	CIFRA_CALC	S_RIVER_TRANSPORT_WORK_X_DAY	DAY	День - партиция (техн.)	string	+			

Скрипты обновления DDL комментариев

attribute_nm	hive_ty			comment	comment	sql_script
				Ctrl+клик чтобы открыть в SQL консоли		
day	string	string	true	День - партиция (техн.)	NotNull | День - партиция (тех	alter table cifra_calc.s_passenger_pa
pmd_type_nm	string	string	true	PK | NotNull | D | Тип СИМ	PK | NotNull | D | Тип СИМ	Select 'Обновление не требуется'
zone_nm	string	string	true	PK | NotNull | D | Зона	PK | NotNull | D | Зона	Select 'Обновление не требуется'
pmd_trip_avg_speed	double	double	true	NotNull | M | Средняя скорост	NotNull | M | Средняя скорости	Select 'Обновление не требуется'
subject_type_nm	string	string	true	PK | NotNull | D | Тематика об	PK | NotNull | D | Тематика обр	Select 'Обновление не требуется'
day	string	string	true	NotNull | День - партиция (те	NotNull | День - партиция (тех	Select 'Обновление не требуется'
appeal_cnt	int	int	true	NotNull | M | Количество случ	NotNull | M | Количество случа	Select 'Обновление не требуется'
esoo_appeal_cnt	int	int	true	NotNull | M | Количество обра	NotNull | M | Количество обра	Select 'Обновление не требуется'
week_num	int	int	true	NotNull | D | Номер отчётной	NotNull | D | Номер отчётной н	Select 'Обновление не требуется'
day	string	string	true	Партиция (техн.)	NotNull | День - партиция (тех	alter table cifra_calc.s_public_transp
day	string	string	true	День - партиция (техн.)	NotNull | День - партиция (тех	alter table cifra_calc.s_public_transp
process_dttm	timestamp	timestamp	true	NotNull | Дата и время добав	NotNull | Дата и время добавл	Select 'Обновление не требуется'
trip_cnt	int	int	true	NotNull | M | Фактическое кол	NotNull | M | Фактическое коли	Select 'Обновление не требуется'
operator_nm	string	string	true	PK | NotNull | D | Перевозчик	PK | NotNull | D | Перевозчик	Select 'Обновление не требуется'
route_num	string	[NULL]	false	PK | NotNull | D | Номер марш	[NULL]	[NULL]
day	string	[NULL]	false	День - партиция (техн.)	[NULL]	[NULL]
division_nm	string	string	true	PK | D | AO	PK | D | AO	Select 'Обновление не требуется'

Очередные успехи

Проверки не для всех таблиц

Устаревание проверок при изменении таблиц

Огромный объём monkey job

Оповещения на почту, которую нельзя подключить к телефону

Логические ошибки в проверках

Наложение выполнения проверок на обновление данных

Задержки в получении результатов проверок

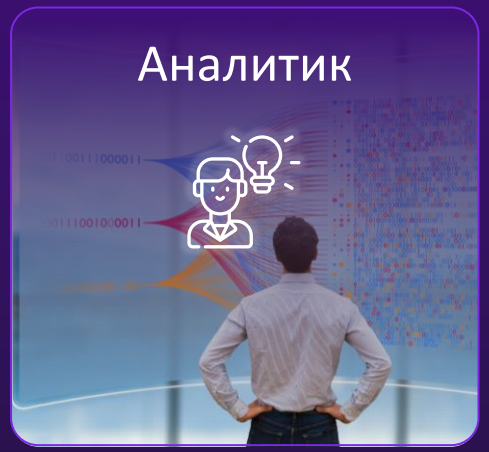
Отдельные сотрудники на мониторинг DQ

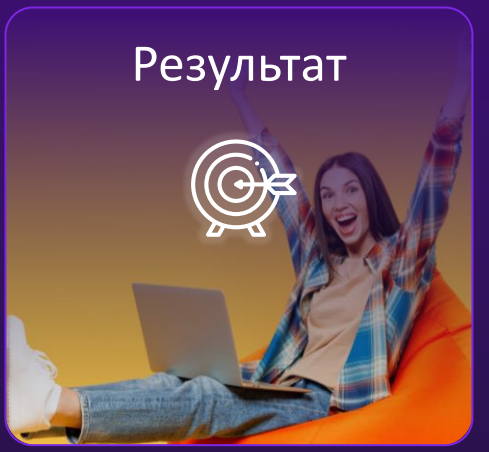
+БОНУС!
ddl таблицы автоматически генерируется сразу после написания документации

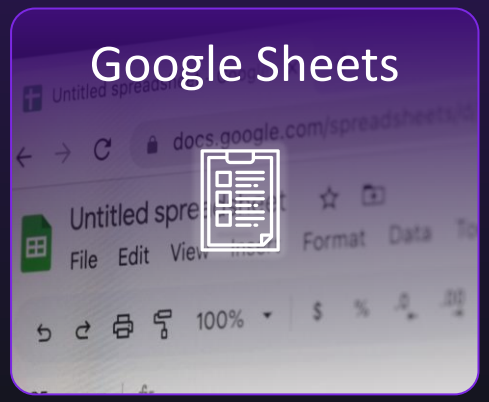
Постоянная доработка бордов

Безопасный транспорт ЦОДД

Оповещения через Telegram

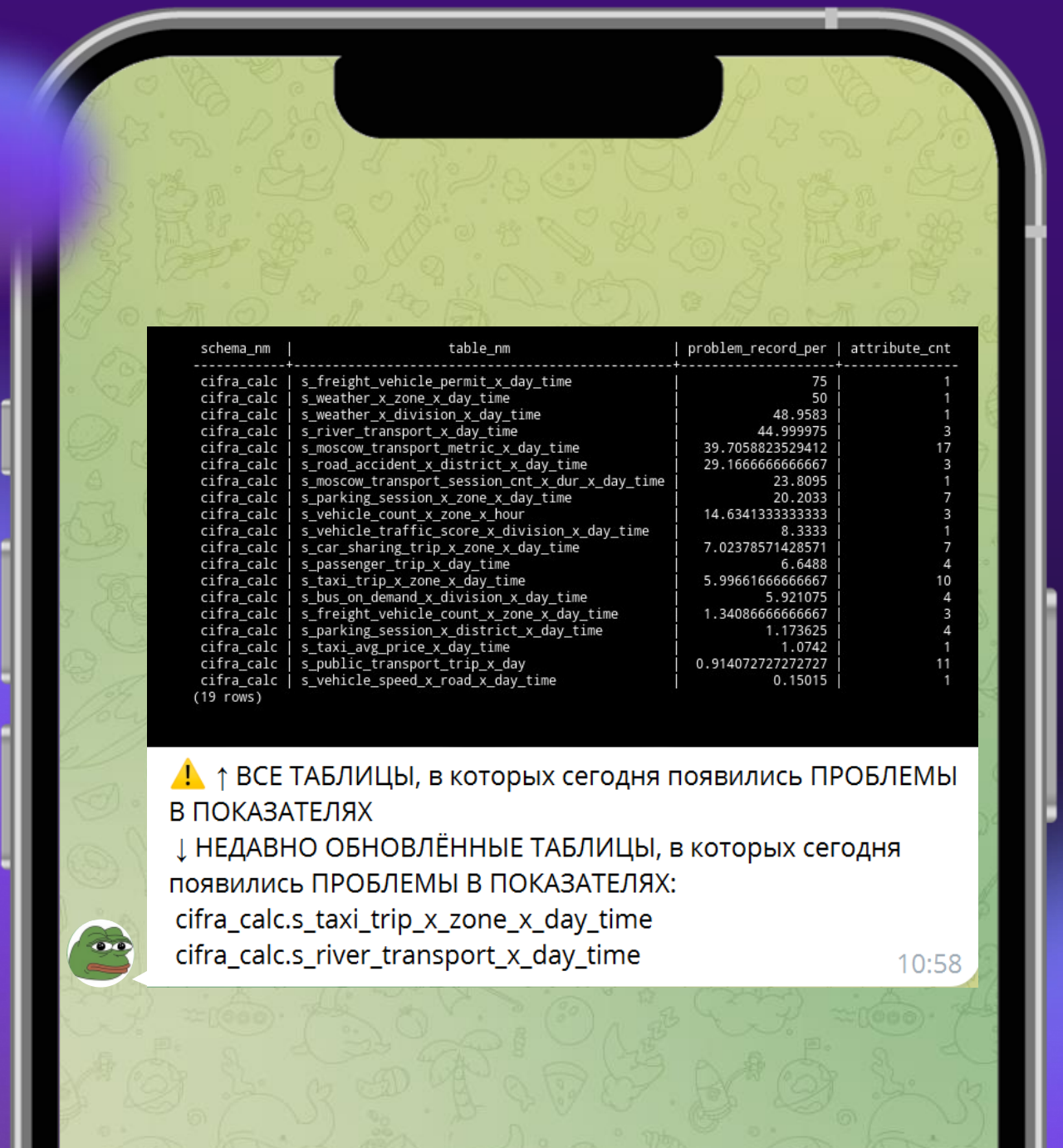






Безопасный транспорт ЦОДД

Telegram bot



schema_nm	table_nm	problem_record_per	attribute_cnt
cifra_calc	s_freight_vehicle_permit_x_day_time	75	1
cifra_calc	s_weather_x_zone_x_day_time	50	1
cifra_calc	s_weather_x_division_x_day_time	48.9583	1
cifra_calc	s_river_transport_x_day_time	44.999975	3
cifra_calc	s_moscow_transport_metric_x_day_time	39.7058823529412	17
cifra_calc	s_road_accident_x_district_x_day_time	29.1666666666667	3
cifra_calc	s_moscow_transport_session_cnt_x_dur_x_day_time	23.8095	1
cifra_calc	s_parking_session_x_zone_x_day_time	20.2033	7
cifra_calc	s_vehicle_count_x_zone_x_hour	14.6341333333333	3
cifra_calc	s_vehicle_traffic_score_x_division_x_day_time	8.3333	1
cifra_calc	s_car_sharing_trip_x_zone_x_day_time	7.02378571428571	7
cifra_calc	s_passenger_trip_x_day_time	6.6488	4
cifra_calc	s_taxi_trip_x_zone_x_day_time	5.99661666666667	10
cifra_calc	s_bus_on_demand_x_division_x_day_time	5.921075	4
cifra_calc	s_freight_vehicle_count_x_zone_x_day_time	1.34086666666667	3
cifra_calc	s_parking_session_x_district_x_day_time	1.173625	4
cifra_calc	s_taxi_avg_price_x_day_time	1.0742	1
cifra_calc	s_public_transport_trip_x_day	0.914072727272727	11
cifra_calc	s_vehicle_speed_x_road_x_day_time	0.15015	1

(19 rows)

⚠️ ↑ ВСЕ ТАБЛИЦЫ, в которых сегодня появились ПРОБЛЕМЫ В ПОКАЗАТЕЛЯХ

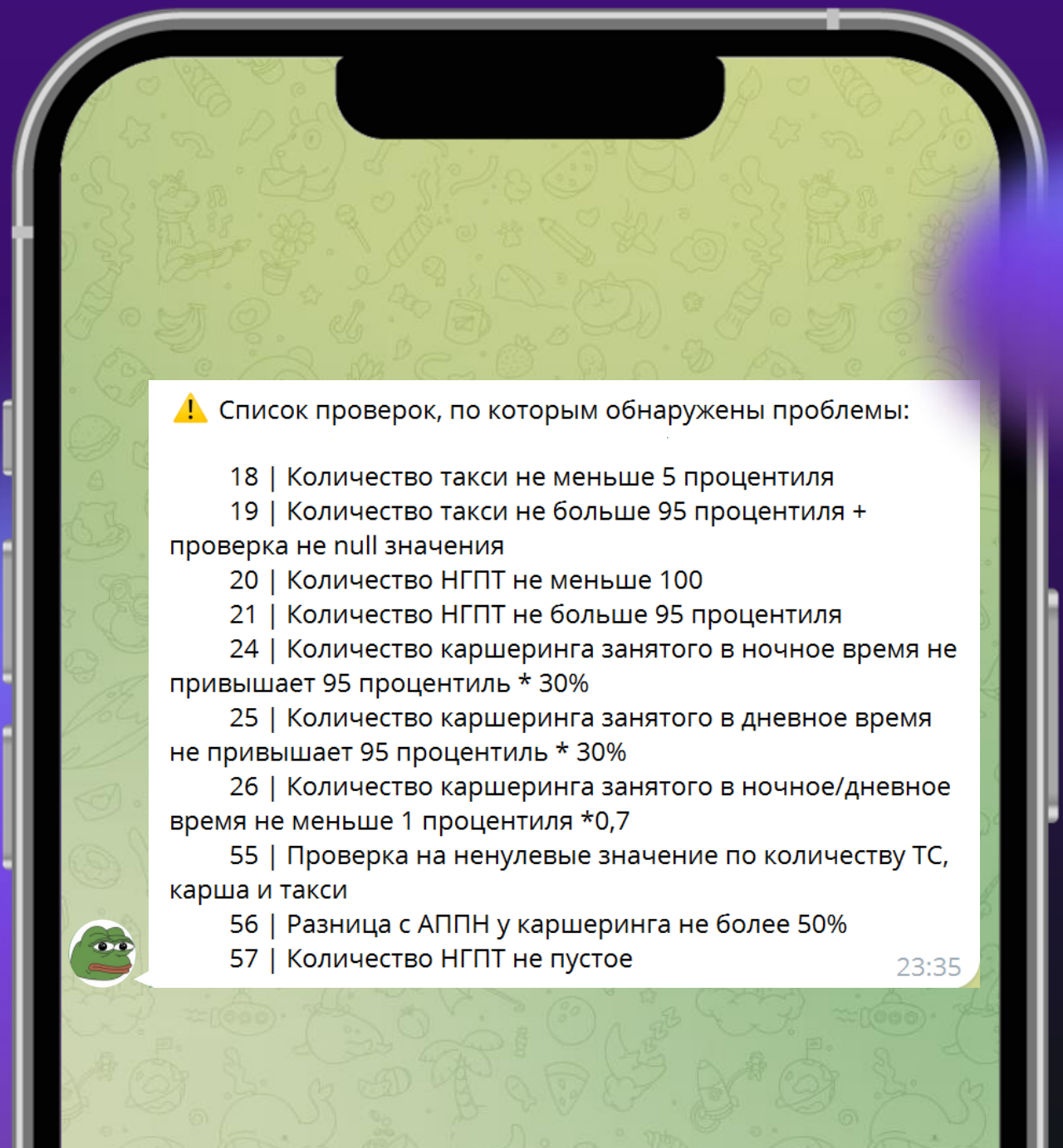
↓ НЕДАВНО ОБНОВЛЁННЫЕ ТАБЛИЦЫ, в которых сегодня появились ПРОБЛЕМЫ В ПОКАЗАТЕЛЯХ:

cifra_calc.s_taxi_trip_x_zone_x_day_time

cifra_calc.s_river_transport_x_day_time

10:58

Telegram bot



Это победа?

Самые важные проверки создаются сами на основе документации

Проверки сами начинают выполняться автоматически после выката в прод дага обновления таблицы

Оповещения по непройденным проверкам сами приходят всем заинтересованным в телефон сразу же, как возникает проблема

Безопасный транспорт ЦОДД

Изменение структуры таблицы 1/2

1

```
CREATE TABLE `cifra_calc.s_public_transport_work_x_day`(
  `dt` date COMMENT 'PK | NotNull | D | Отчётная дата',
  `working_day_flg` int COMMENT 'NotNull | D | Признак рабочего дня',
  `week_num` int COMMENT 'NotNull | D | Номер отчётной недели',
  `vehicle_cnt` int COMMENT 'NotNull | M | Фактическое количество ТС (выпуск)',
  `completed_vehicle_percent` double COMMENT 'NotNull | M | Процент выполнения выпуска',
  `trip_cnt` int COMMENT 'NotNull | M | Фактическое количество рейсов',
  `plan_trip_cnt` int COMMENT 'NotNull | M | Плановое количество рейсов',
  `plan_driver_cnt` int COMMENT 'NotNull | M | Плановое количество водителей',
  `completed_trip_percent` double COMMENT 'NotNull | M | Процент выполнения рейсов',
  `process_dttm` timestamp COMMENT 'NotNull | Дата и время добавления записи в таблицу (техн.)')
PARTITIONED BY (
  `day` string COMMENT 'NotNull | День - партиция (техн.)')
```

2

Код схемы	Код таблицы	Код атрибута	Описание атрибута	Тип данных в Hive	Обязательное	Вхождение в первичный ключ	Измерение Dimension	Показатель Measure
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	DT	Отчётная дата	date	+	+	+	
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	WORKING_DAY_FLG	Признак рабочего дня	int	+		+	
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	WEEK_NUM	Номер отчётной недели	int	+		+	
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	OPERATOR_NM	Перевозчик	string	+	+	+	
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	VEHICLE_CNT	Фактическое количество ТС (выпуск)	int	+			+
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	PLAN_VEHICLE_CNT	Плановое количество ТС (выпуск)	int	+			+
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	COMPLETED_VEHICLE_PERCENT	Процент выполнения выпуска	double	+			+
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	TRIP_CNT	Фактическое количество рейсов	int	+			+
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	PLAN_TRIP_CNT	Плановое количество рейсов	int	+			+
~~CIFRA_CALC~~	~~S_PUBLIC_TRANSPORT_WORK_X_DAY~~	~~PLAN_DRIVER_CNT~~	~~Плановое количество водителей~~	~~int~~	~~+~~			~~+~~
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	COMPLETED_TRIP_PERCENT	Процент выполнения рейсов	double	+			+
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	PROCESS_DTTM	Дата и время добавления записи в таблицу (техн.)	timestamp	+			
CIFRA_CALC	S_PUBLIC_TRANSPORT_WORK_X_DAY	DAY	День - партиция (техн.)	string	+			

Изменение структуры таблицы 2/2

3

4

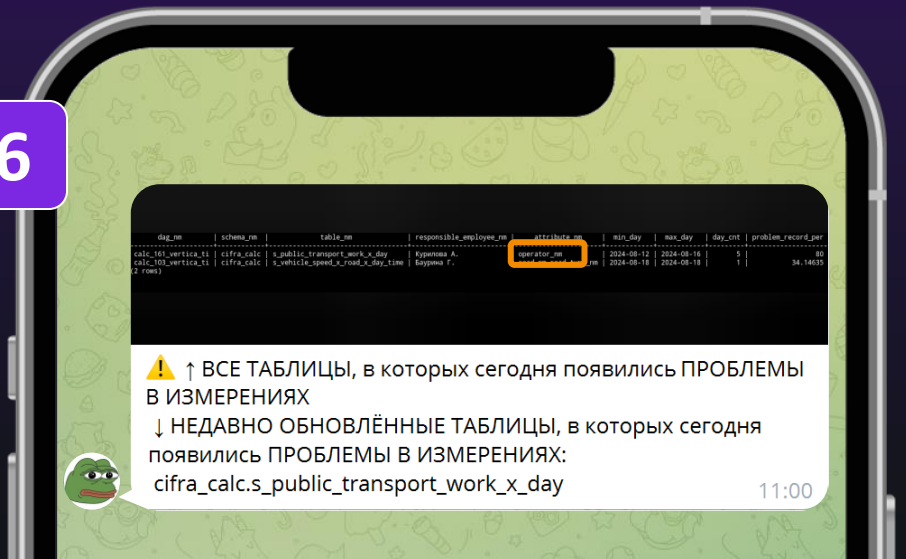
5

6

```
select * from data_quality.metrics where  Введите SQL выражение чтобы отфильтровать результаты
```

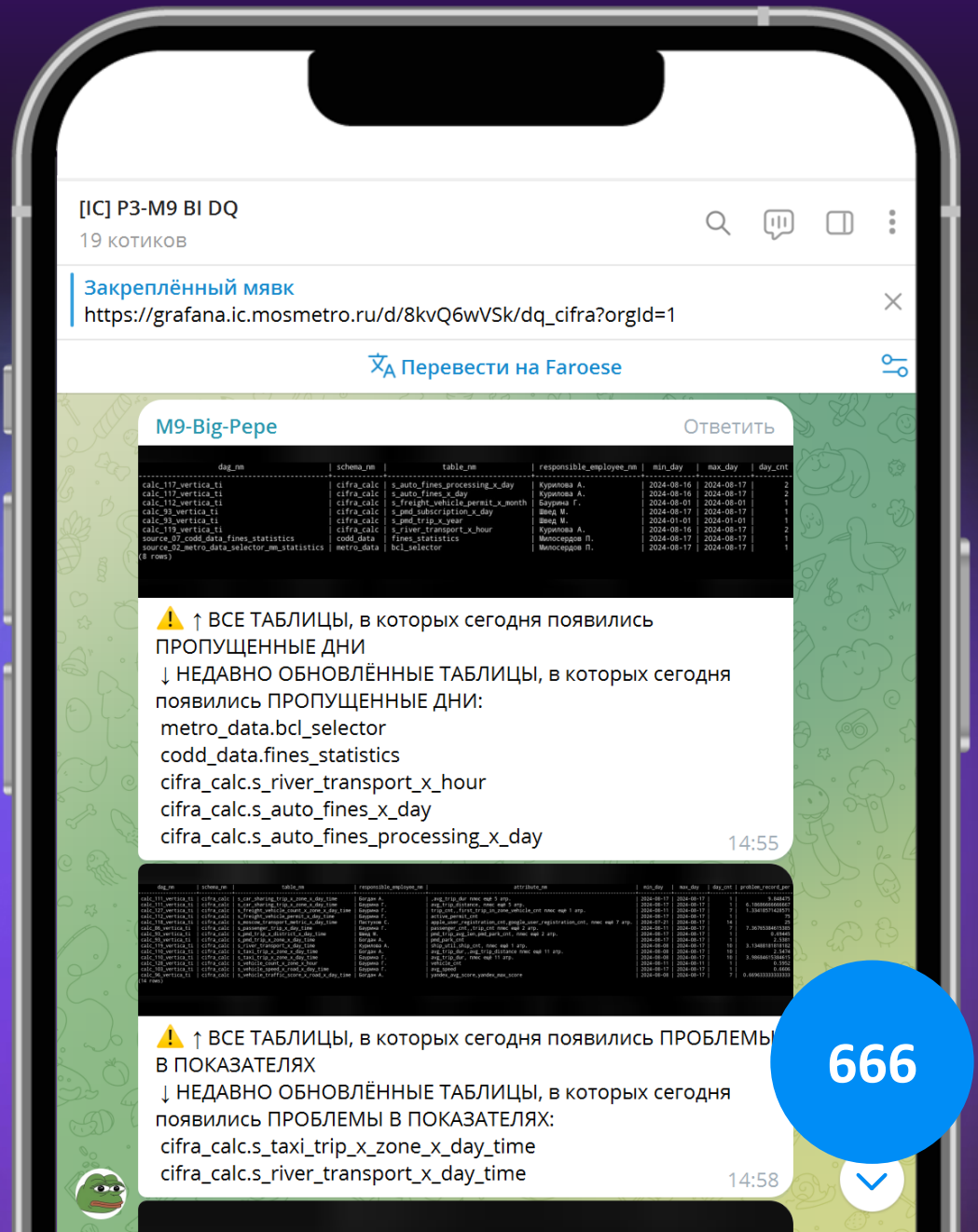
	attribute_nm	attribute_desc	metric_type_cd	sql_script	metric_nm
7	operator_nm	Отчётная дата; Перевозчик	doubles	SELECT CASE WHEN COUNT(*	Проверка на отсутствие дублей по кл
8	dt; working_day_flg; week_num operator_nm	Отчётная дата; Признак раб	null_attributes	SELECT CASE WHEN COUNT(*	Проверка на отсутствие пустых значе
9	operator_nm	Перевозчик	dimension_domain	WITH t1 AS (SELECT 1 AS chec	Проверка на соответствие значения р
10	operator_nm	Перевозчик	dimension_domain	SELECT CASE WHEN COUNT(*	Проверка на соответствие значения р
11	operator_nm	Перевозчик	lost_keys	SELECT CASE WHEN COUNT(*	Проверка на отсутствие потерянных к
12	operator_nm	Перевозчик	extra_keys	SELECT CASE WHEN COUNT(*	Проверка на отсутствие лишних ключ
13	plan_trip_cnt	Плановое количество рейсов	measure_domain	SELECT CASE WHEN COUNT(*	Проверка на соответствие значения п
14	plan_trip_cnt	Плановое количество рейсов	outliers	WITH t1 AS (SELECT COALESC	Проверка на отсутствие выбросов в п
15	plan_vehicle_cnt	Плановое количество ТС (вы	measure_domain	SELECT CASE WHEN COUNT(*	Проверка на соответствие значения п
16	plan_vehicle_cnt	Плановое количество ТС (вы	outliers	WITH t1 AS (SELECT COALESC	Проверка на отсутствие выбросов в п
17	trip_cnt	Фактическое количество рейс	measure_domain	SELECT CASE WHEN COUNT(*	Проверка на соответствие значения п
18	trip_cnt	Фактическое количество рейс	outliers	WITH t1 AS (SELECT COALESC	Проверка на отсутствие выбросов в п
19	vehicle_cnt	Фактическое количество ТС (в	outliers	WITH t1 AS (SELECT COALESC	Проверка на отсутствие выбросов в п
20	vehicle_cnt	Фактическое количество ТС (в	measure_domain	SELECT CASE WHEN COUNT(*	Проверка на соответствие значения п
21	week_num	Номер отчётной недели	dimension_domain	SELECT 100	Проверка на соответствие значения р

```
CREATE TABLE `cifra_calc.s_public_transport_work_x_day`(
  `dt` date COMMENT 'PK | NotNull | D | Отчётная дата',
  `working_day_flg` int COMMENT 'NotNull | D | Признак рабочего дня',
  `week_num` int COMMENT 'NotNull | D | Номер отчётной недели',
  `operator_nm` string COMMENT 'PK | NotNull | D | Перевозчик',
  `vehicle_cnt` int COMMENT 'NotNull | M | Фактическое количество ТС (выпуск)',
  `plan_vehicle_cnt` int COMMENT 'NotNull | M | Плановое количество ТС (выпуск)',
  `completed_vehicle_percent` double COMMENT 'NotNull | M | Процент выполнения выпуска',
  `trip_cnt` int COMMENT 'NotNull | M | Фактическое количество рейсов',
  `plan_trip_cnt` int COMMENT 'NotNull | M | Плановое количество рейсов',
  `completed_trip_percent` double COMMENT 'NotNull | M | Процент выполнения рейсов',
  `process_dttm` timestamp COMMENT 'NotNull | Дата и время добавления записи в таблицу (техн.)')
PARTITIONED BY (
  `day` string COMMENT 'NotNull | День - партиция (техн.)')
```



Безопасный транспорт ЦОДД

Слишком много сообщений от бота

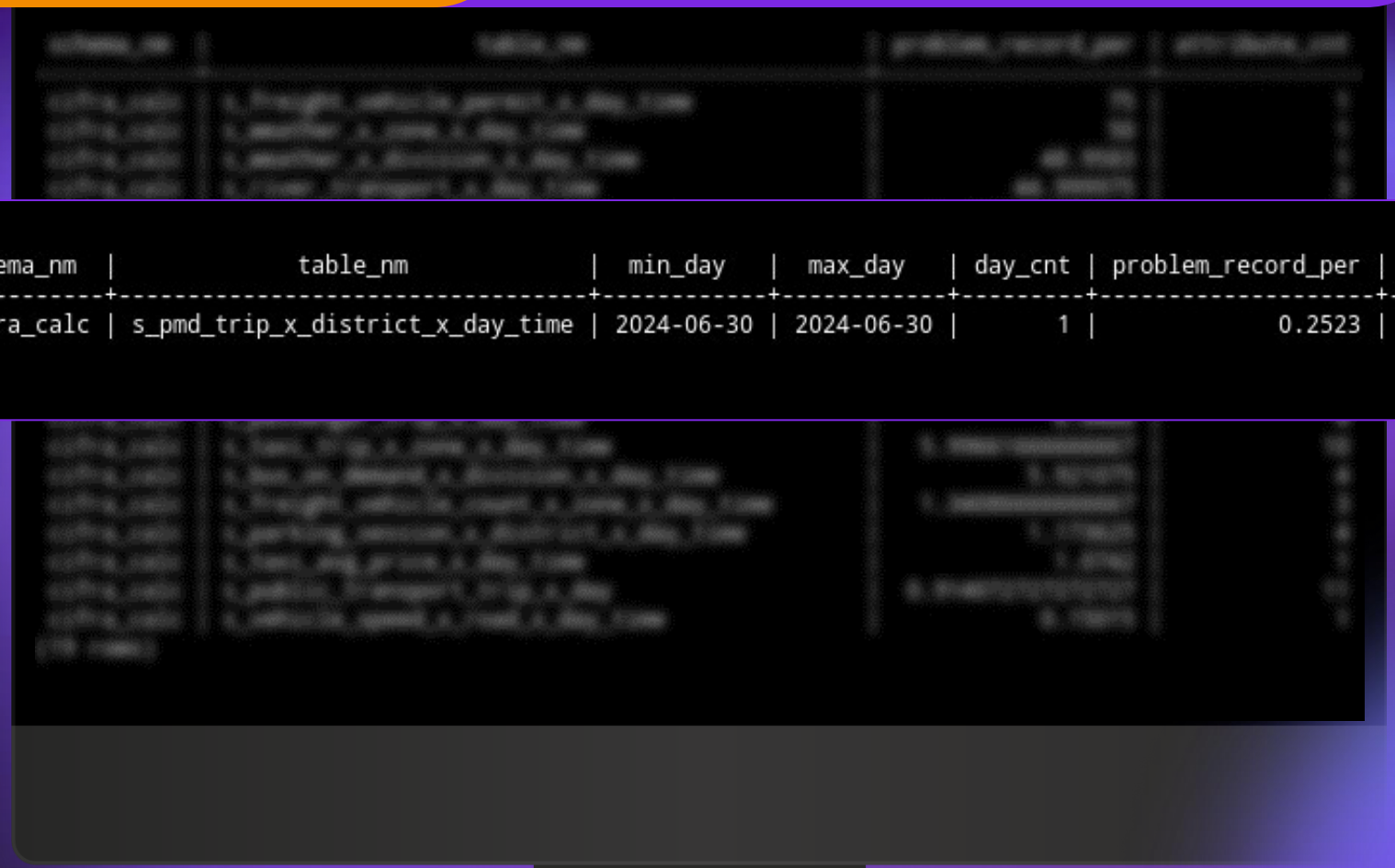


Безопасный транспорт ЦОДД

Telegram bot

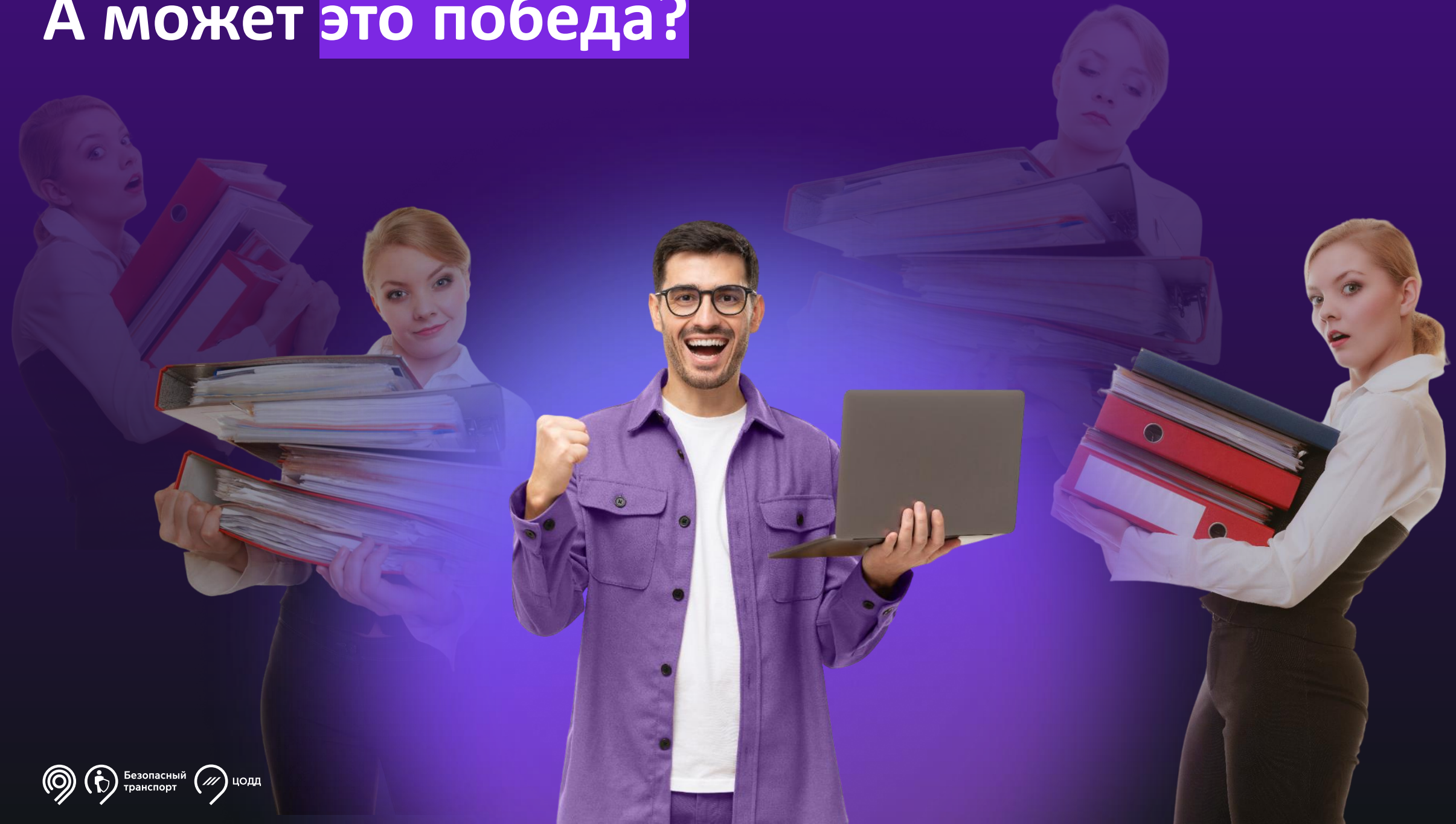
@galina_baurina

Таблицы, по которым есть дубли



```
       dag_nm       | schema_nm  |             table_nm              |  min_day   |  max_day   | day_cnt | problem_record_per | responsible_employee_nm
--------------------+------------+-----------------------------------+------------+------------+---------+--------------------+-------------------------
 calc_93_vertica_ti | cifra_calc | s_pmd_trip_x_district_x_day_time | 2024-06-30 | 2024-06-30 |       1 |             0.2523 | Баурина Г.
(1 row)
```

А может это победа?



Безопасный транспорт ЦОДД

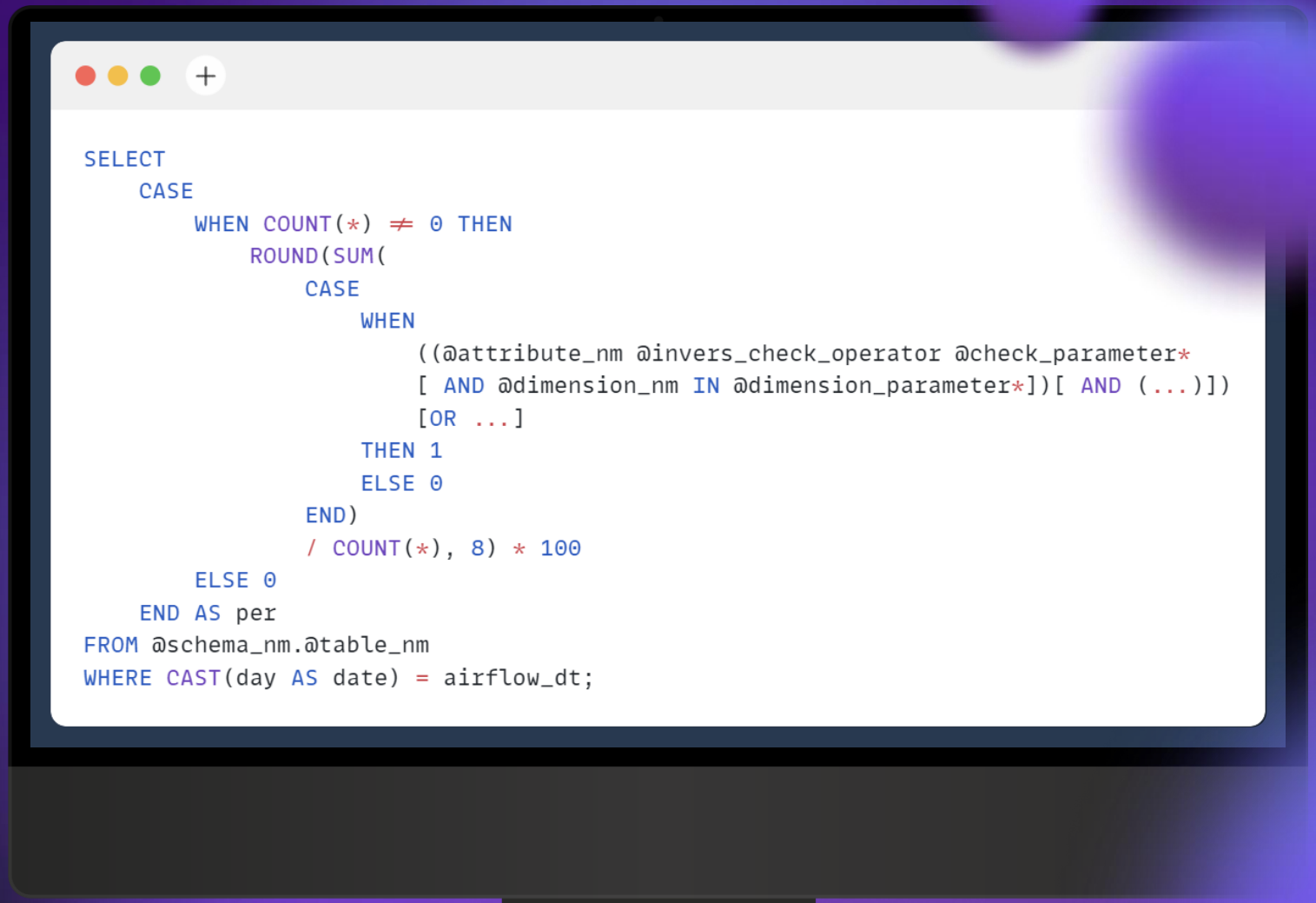
Ручные проверки

schedule	data_base_nm	schema_nm	table_nm	metric_type_cd	metric_num	responsible_employee_nm	metric_desc	expected_result	sql_metric_script	sql_period_frm_dt
minute 5	vertica_ti	yandex_data	jams	count	metric_nun	Баурина Г.	Нет балла Яндекса за последние 10 минут	0	with t_isx as (select response_dttm, sysdate - response_dttm from yandex_data.jams order by response_dttm desc limit 1) select case when delta > '00:10:00' then 1 else (from t_isx	select airflow_(
realtime	vertica_ti	transport_inde…	s_mt_city	count	metric_nun	Мадумаров А.	Нет УО для магистрали	0	select case when service_level_highway is null t from transport_index_calc.s_mt_city where five_min_ts = (select max(five_min_ts) fro	select airflow_(
realtime	vertica_ti	transport_inde…	s_mt_city	logic	metric_nun	Баурина Г.	УО для магистрали ниже/выше по городу на 2 уровня	0	select case when service_level='A' and service_l when service_level='B' and servic when service_level='C' and servic when service_level='D' and servic when service_level='E' and servic	select airflow_(
minute 5	vertica_ti	transport_inde…	s_mt_pivot_vw	logic	metric_nun	Мадумаров А.	Балл яндекса и цодд отличается больше чем на 2	0	select case when abs(score-yandex_score)>=2 from transport_index_calc.s_mt_pivot_vw	select airflow_(
minute 5	vertica_ti	transport_inde…	s_mt_pivot_vw	count	metric_nun	Мадумаров А.	Количество ТС нулевое или пустое	0	select case when regno_cnt is null or regno_cnt= from transport_index_calc.s_mt_pivot_vw	select airflow_(
minute 5	vertica_ti	transport_inde…	s_mt_service_level_highway_cnt	logic	metric_nun	Мадумаров А.	По округу наиболее частотный УО отличается от города больше чем на 2	0	with okrug_level as (select count_level, service_level, ten_min_ts from transport_index_calc.s_mt_service_level_hi where detail ='Округ' and count_level = (select max(count_level) from where detail ='Округ' limit 1) limit 1) , city_level as (select five_min_ts , service_level from transport_index_calc.s_mt_city	select airflow_(
off	vertica_ti	transport_inde…	s_mt_service_level_highway_cnt	count	metric_nun	Баурина Г.	Нет УО по округу	0	with okrug_level as (select count_level, service_level, ten_min_ts from transport_index_calc.s_mt_service_level_hi where detail ='Округ' and count_level = (select max(count_level) from where detail ='Округ' limit 1)	select airflow_(

Параметры автоматических проверок

schema_nm	table_nm	attribute_nm	metric_type_cd	check_group_num	check_operator	check_parameter	dimension_nm	dimension_operator	dimension_parameter
metro_data	bcl_selector	month	domain	1	<	1			
metro_data	bcl_selector	month	domain	2	>	12			
metro_data	bcl_selector	pass_cnt	domain	1	<	0			
metro_data	bcl_selector	pass_cnt	domain	2	>	500000	working_day_flg	IN	Нет, 5 дней фри, вых
metro_data	bcl_selector	pass_cnt	domain	3	>	700000	working_day_flg	IN	Нет;Между майскими;Нет, 5 д
metro_data	bcl_selector	pass_cnt	domain	4	>	800000	working_day_flg	IN	Да
metro_data	bcl_selector	pass_cnt_transfers	domain	1	<	0			
metro_data	bcl_selector	pass_cnt_transfers	domain	2	>	500000	working_day_flg	IN	Нет, 5 дней фри, вых
metro_data	bcl_selector	pass_cnt_transfers	domain	3	>	800000	working_day_flg	IN	Нет;Между майскими;Нет, 5 д
metro_data	bcl_selector	pass_cnt_transfers	domain	4	>	1000000	working_day_flg	IN	Да
metro_data	bcl_selector	pass_cnt_with_transfers	domain	1	<	0			
metro_data	bcl_selector	pass_cnt_with_transfers	domain	2	>	1000000	working_day_flg	IN	Нет, 5 дней фри, вых
metro_data	bcl_selector	pass_cnt_with_transfers	domain	3	>	1500000	working_day_flg	IN	Нет;Между майскими;Нет, 5 д
metro_data	bcl_selector	pass_cnt_with_transfers	domain	4	>	1800000	working_day_flg	IN	Да
metro_data	bcl_selector	working_day_flg	domain	1	NOT IN	Да;Нет;Между майскими;Нет, 5 дне			
metro_data	bcl_selector		count	1					
metro_data	bcl_selector		doubles	1					
metro_data	bcl_selector		null_attributes	1					
metro_data	bus_station_pass_statistics	pass_cnt	domain	1	<	0			
metro_data	bus_station_pass_statistics	pass_cnt	domain	2	>	40000			
metro_data	bus_station_pass_statistics		count	1					
metro_data	bus_station_pass_statistics		doubles	1					
metro_data	bus_station_pass_statistics		null_attributes	1					
metro_data	ctu_selector_daily	direction	domain	1	NOT IN	Калужское;Павелецкое;Савеловско			
metro_data	ctu_selector_daily	line	domain	1	NOT IN	МЦД-1;МЦД-2;МЦД-3;МЦД-4;МЦД			
metro_data	ctu_selector_daily	passflow	domain	1	<	0			
metro_data	ctu_selector_daily	passflow	domain	2	>	250000	direction	IN	Калужское;Павелецкое;Савело

Шаблон проверки Domain



```
SELECT
    CASE
        WHEN COUNT(*) =/= 0 THEN
            ROUND(SUM(
                CASE
                    WHEN
                        ((@attribute_nm @invers_check_operator @check_parameter*
                        [ AND @dimension_nm IN @dimension_parameter*])[ AND (...)])
                        [OR ...]
                    THEN 1
                    ELSE 0
                END)
                / COUNT(*), 8) * 100
        ELSE 0
    END AS per
FROM @schema_nm.@table_nm
WHERE CAST(day AS date) = airflow_dt;
```

Пример параметров проверки Domain без зависимости от дат

table_nm	attribute_nm	metric_type_cd	check_group_num	check_operator	check_parameter	dimension_nm	dimension_operator	dimension_parameter	responsible_employee_nm
s_airport_trip_x_day	passenger_cnt	domain	1	<	1000	flight_type_nm	IN	Внутренние рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	1	<	1000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	2	>	120000	flight_type_nm	IN	Внутренние рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	2	>	120000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	3	<	10	flight_type_nm	IN	Международные рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	3	<	10	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	4	>	120000	flight_type_nm	IN	Международные рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	4	>	120000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.

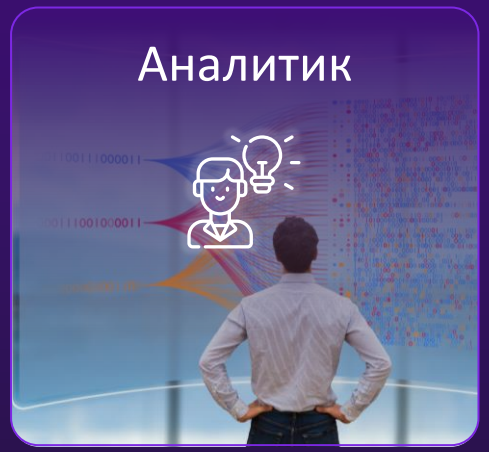
Пример параметров проверки Domain с зависимостью от дат

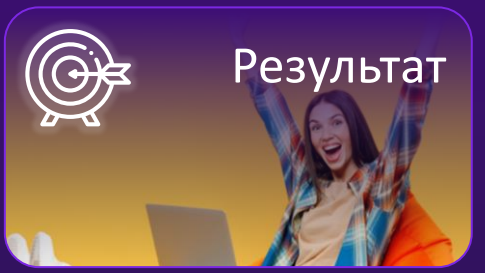
table_nm	attribute_nm	metric_type_cd	check_group_num	check_operator	check_parameter	dimension_nm	dimension_operator	dimension_parameter	responsible_employee_nm
s_airport_trip_x_day	passenger_cnt	domain	1	<	1000	flight_type_nm	IN	Внутренние рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	1	<	1000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	1	<	1000	day	<	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	2	>	120000	flight_type_nm	IN	Внутренние рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	2	>	120000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	2	>	120000	day	<	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	3	<	10	flight_type_nm	IN	Международные рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	3	<	10	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	3	<	10	day	<	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	4	>	120000	flight_type_nm	IN	Международные рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	4	>	120000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	4	>	120000	day	<	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	5	<	100	flight_type_nm	IN	Внутренние рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	5	<	100	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	5	<	100	day	>=	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	6	>	12000	flight_type_nm	IN	Внутренние рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	6	>	12000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	6	>	12000	day	>=	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	7	<	10	flight_type_nm	IN	Международные рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	7	<	10	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	7	<	10	day	>=	2020-03-01	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	8	>	12000	flight_type_nm	IN	Международные рейсы	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	8	>	12000	direction_nm	IN	Прибывшие;Убывшие	Баурина Г.
s_airport_trip_x_day	passenger_cnt	domain	8	>	12000	day	>=	2020-03-01	Баурина Г.

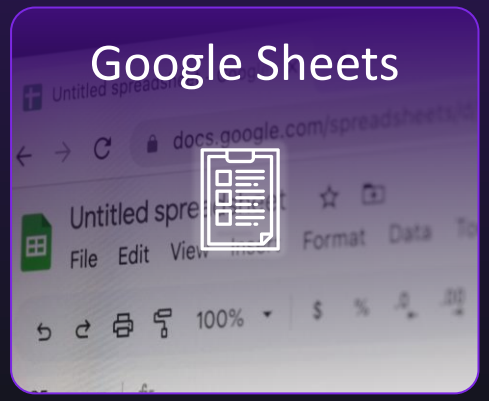
Пример скрипта проверки Domain с зависимостью от дат

```sql
SELECT
    CASE
        WHEN COUNT(*) != 0 THEN ROUND(SUM(CASE WHEN
        ((passenger_cnt < 1000 AND direction_nm IN ('Прибывшие', 'Убывшие') AND flight_type_nm IN ('Внутренние рейсы') AND day < '2020-03-01' ))
        OR ((passenger_cnt > 120000 AND direction_nm IN ('Прибывшие', 'Убывшие') AND flight_type_nm IN ('Внутренние рейсы') AND day < '2020-03-01'))
        OR ((passenger_cnt < 10 AND flight_type_nm IN ('Международные рейсы') AND direction_nm IN ('Прибывшие', 'Убывшие') AND day < '2020-03-01'))
        OR ((passenger_cnt > 120000 AND direction_nm IN ('Прибывшие', 'Убывшие') AND flight_type_nm IN ('Международные рейсы') AND day < '2020-03-01'))

        OR ((passenger_cnt < 100 AND direction_nm IN ('Прибывшие', 'Убывшие') AND flight_type_nm IN ('Внутренние рейсы') AND day >= '2020-03-01' ))
        OR ((passenger_cnt > 12000 AND direction_nm IN ('Прибывшие', 'Убывшие') AND flight_type_nm IN ('Внутренние рейсы') AND day >= '2020-03-01'))
        OR ((passenger_cnt < 10 AND flight_type_nm IN ('Международные рейсы') AND direction_nm IN ('Прибывшие', 'Убывшие') AND day >= '2020-03-01'))
        OR ((passenger_cnt > 12000 AND direction_nm IN ('Прибывшие', 'Убывшие') AND flight_type_nm IN ('Международные рейсы') AND day >= '2020-03-01'))

        THEN 1 ELSE 0 END) / COUNT(*),8) * 100
        ELSE 0
    END AS per
FROM
    cifra_calc.s_airport_trip_x_day
WHERE
    CAST(day AS date) = airflow_dt
```

Пересчёт старых результатов проверок

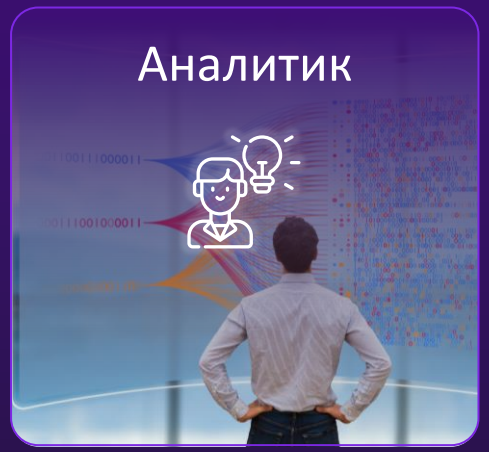


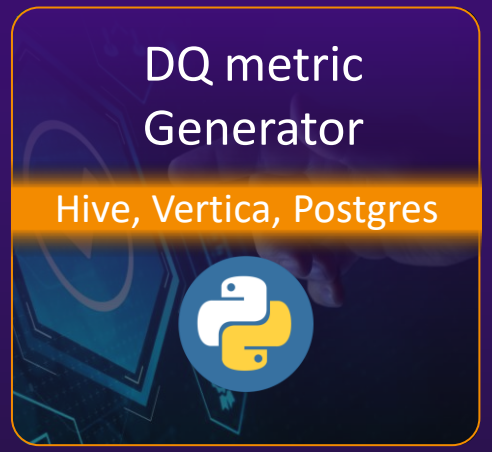


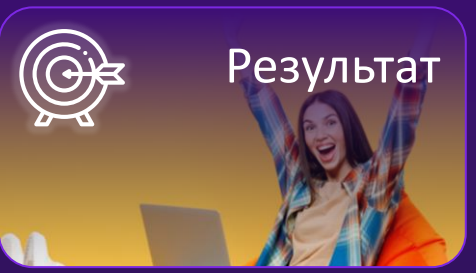


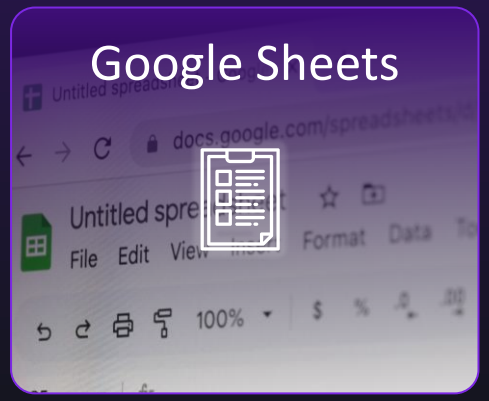
Безопасный транспорт ЦОДД

Добавление SQL-диалектов

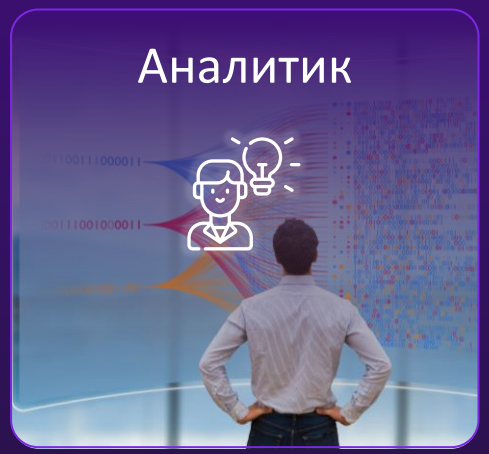


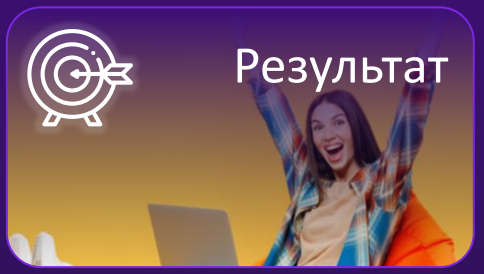


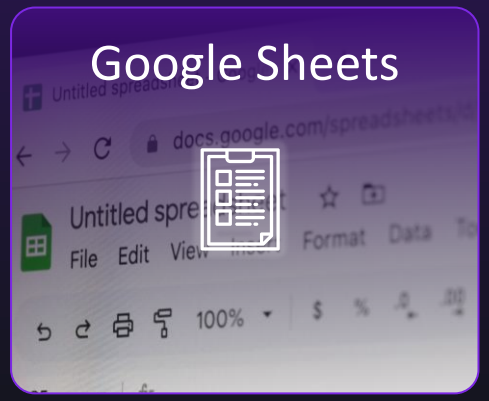




Зависимости между источниками и витринами

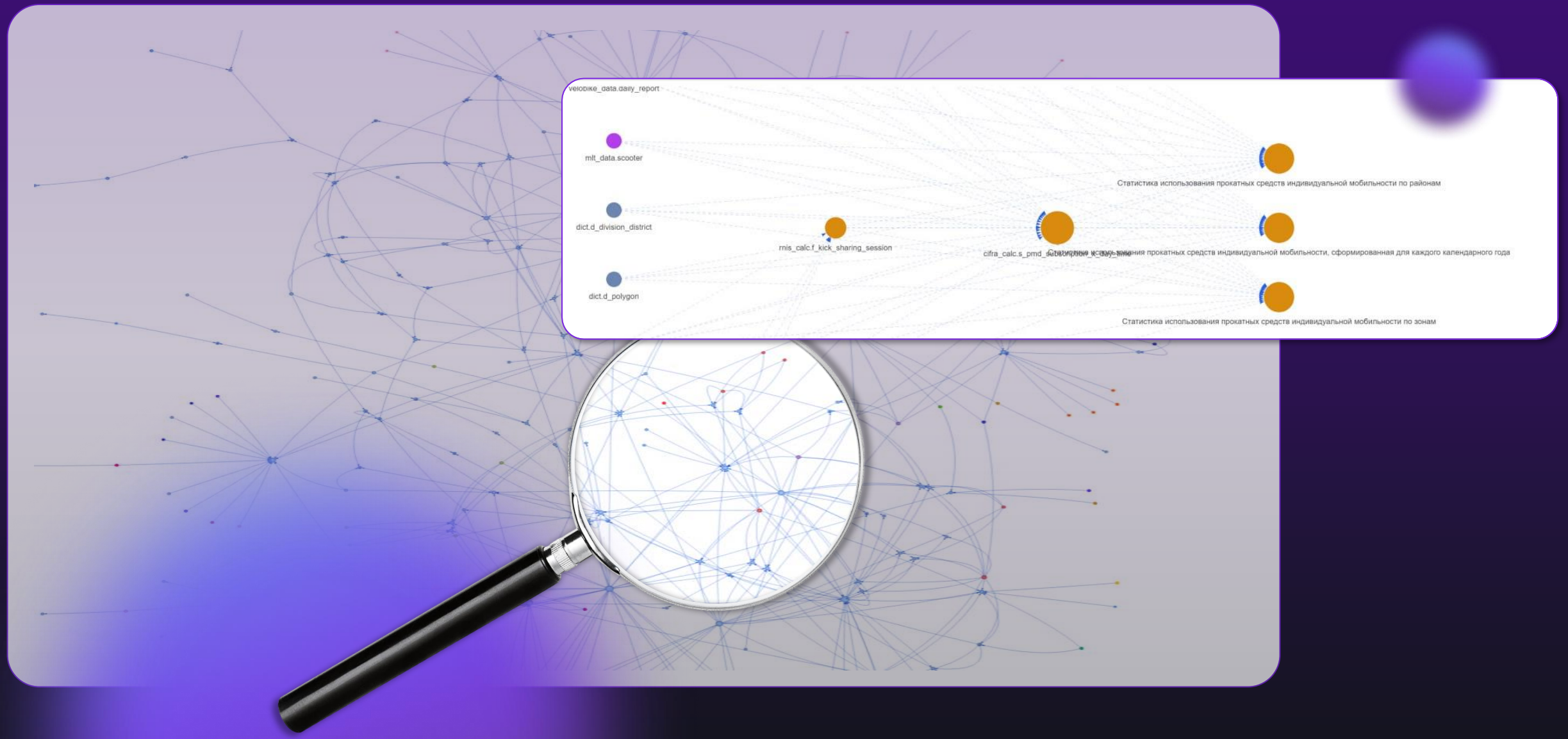




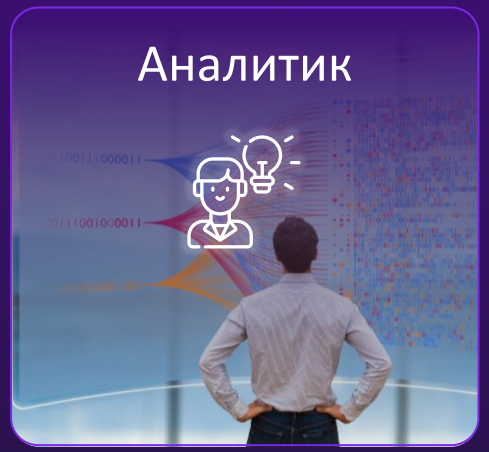


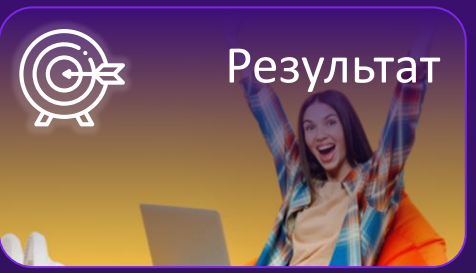
Безопасный транспорт ЦОДД

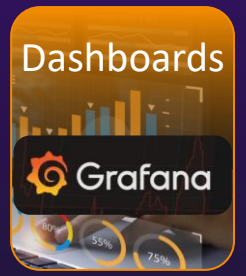
Дерево зависимостей

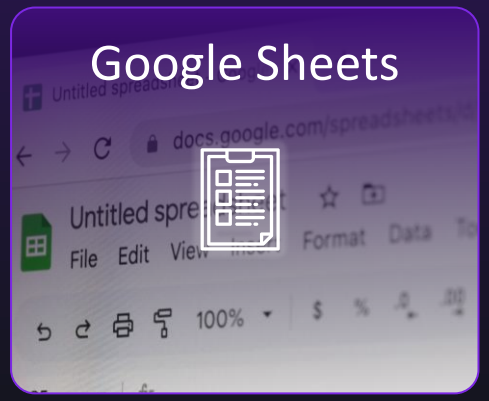


Аналитика ретроспективы



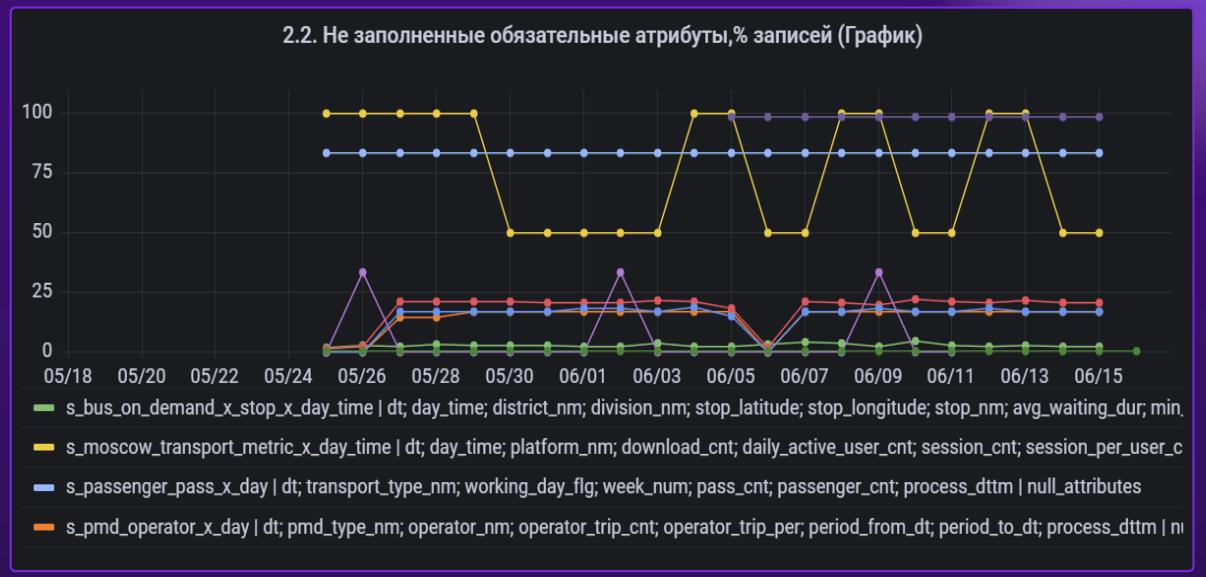


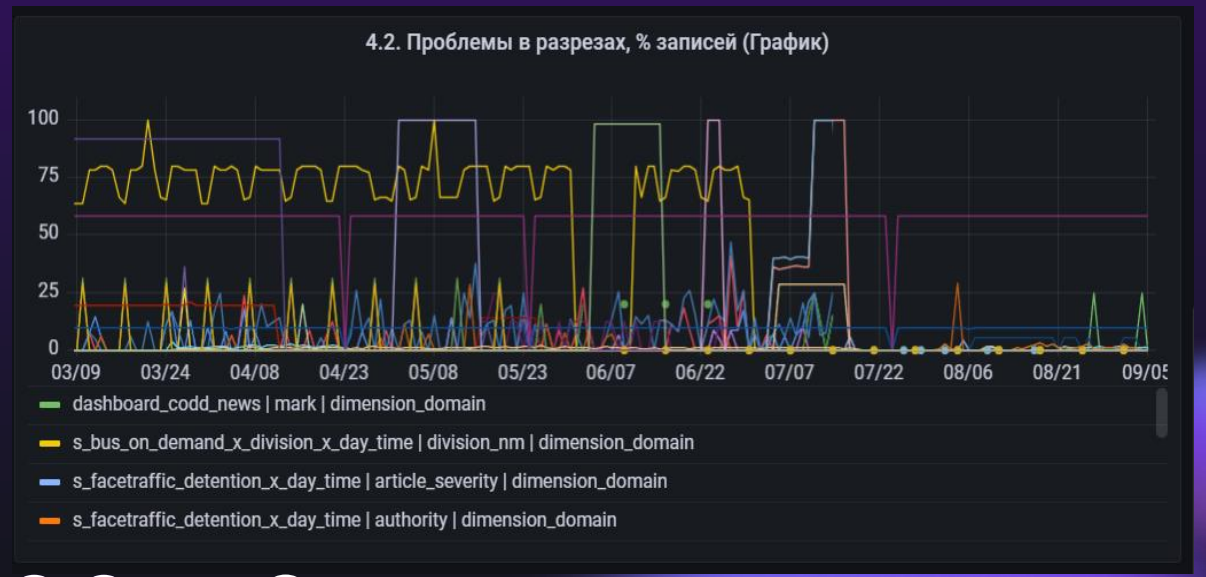


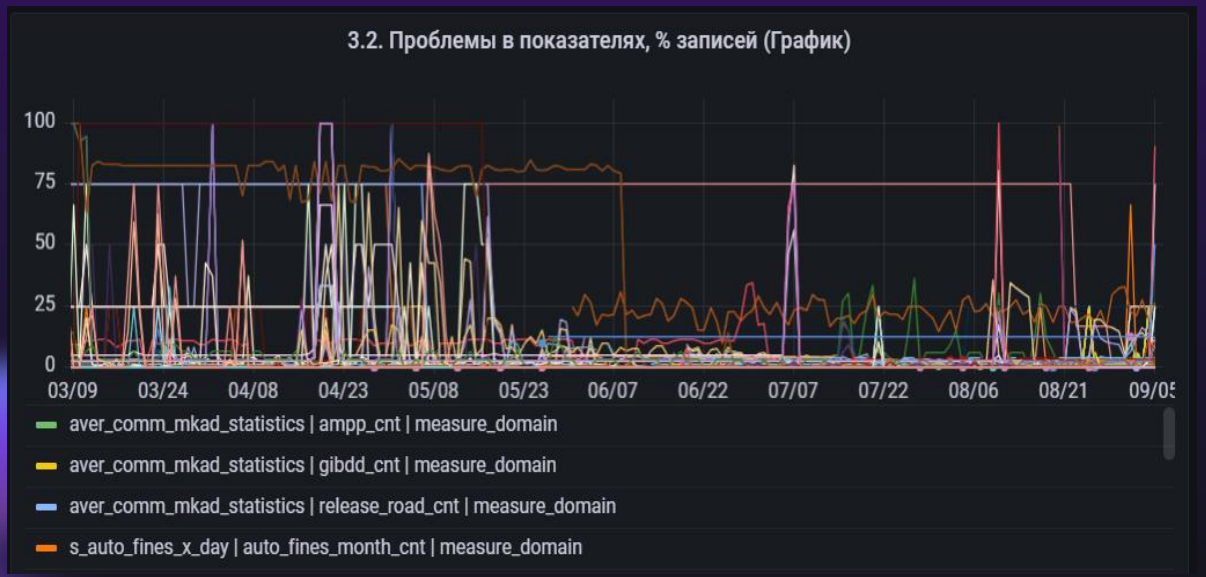


Безопасный транспорт ЦОДД

Grafana





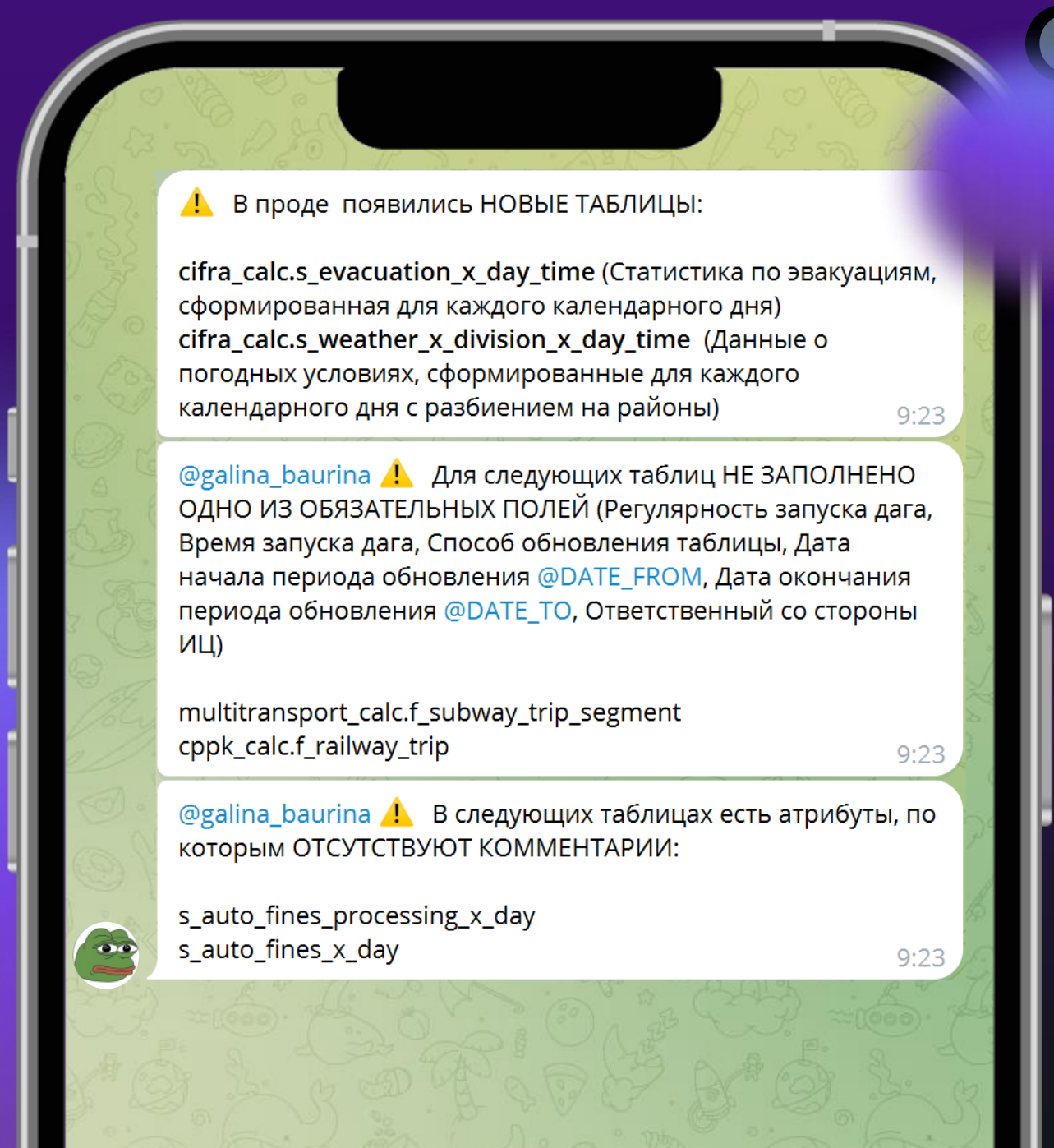


Безопасный транспорт ЦОДД

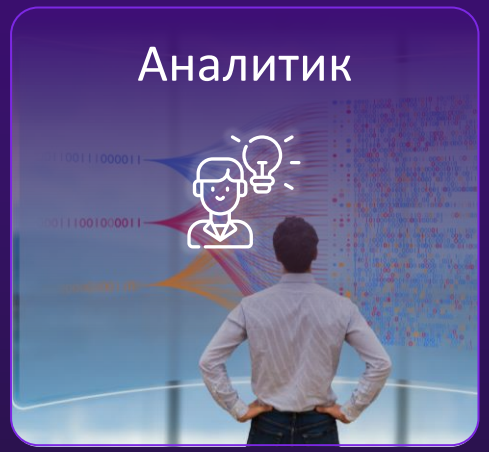
Настройка Telegram-ботов

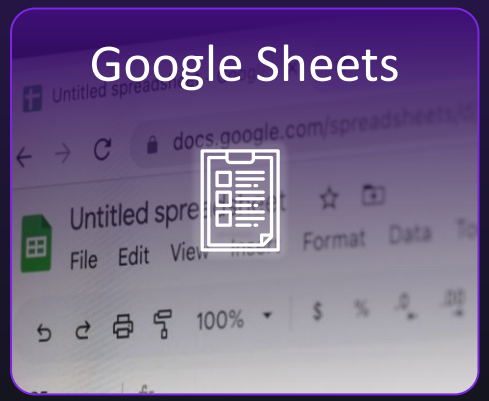
```
# BI DQ Основное печатное сообщение вида 6
MSG[1,6]='↑ ВСЕ ТАБЛИЦЫ, в которых сегодня ИЗМЕНИЛСЯ НАБОР ИЗМЕРЕНИЙ по сравнению с прошедшей неделей'
# BI DQ Основной отчёт, выводимый в картинку вида 6
RRQ[1,6]="SELECT dag_nm, schema_nm, table_nm, responsible_employee_nm, min(day) AS min_day, max(day) AS max_day,
# BI DQ Список телеграмм-ников ответственных вида 6
URQ[1,6]="SELECT listagg(distinct responsible_employee_tg_nick) FROM data_quality.bi_results WHERE process_dttm >
# BI DQ Сокращённый отчёт, выводимый в печатное сообщение вида 6
PLR[1,6]="SELECT * FROM (SELECT '↓ НЕДАВНО ОБНОВЛЁННЫЕ ТАБЛИЦЫ, в которых сегодня ИЗМЕНИЛСЯ НАБОР ИЗМЕРЕНИЙ по ср

# BI DQ Основное печатное сообщение вида 7
MSG[1,7]='↑ ВСЕ ТАБЛИЦЫ, в которых сегодня появились ПРОБЛЕМЫ ПО ПРОЧИМ ПРОВЕРКАМ, СОЗДАННЫМ ВРУЧНУЮ'
# BI DQ Основной отчёт, выводимый в картинку вида 7
RRQ[1,7]="SELECT dag_nm, schema_nm, table_nm, responsible_employee_nm, min(day) AS min_day, max(day) AS max_day,
# BI DQ Список телеграмм-ников ответственных вида 7
URQ[1,7]="SELECT listagg(distinct responsible_employee_tg_nick) FROM data_quality.bi_results WHERE process_dttm >
# BI DQ Сокращённый отчёт, выводимый в печатное сообщение вида 7
PLR[1,7]="SELECT * FROM (SELECT '↓ НЕДАВНО ОБНОВЛЁННЫЕ ТАБЛИЦЫ, в которых сегодня появились ПРОБЛЕМЫ ПО ПРОЧИМ ПР

# BI DQ Основное печатное сообщение вида 8
MSG[1,8]='↑ ВСЕ ТАБЛИЦЫ, в которых не по всем показателям/измерениям созданы ПРАВИЛА ПРОВЕРКИ НА ДИАПАЗОНЫ/ДОМЕНЫ
# BI DQ Основной отчёт, выводимый в картинку вида 8
RRQ[1,8]="SELECT schema_nm, table_nm, responsible_employee_nm, count(*) AS metric_cnt FROM data_quality.bi_metric
# BI DQ Список телеграмм-ников ответственных вида 8
URQ[1,8]="SELECT listagg(distinct responsible_employee_tg_nick) FROM data_quality.bi_metrics WHERE sql_metric_scr
# BI DQ Сокращённый отчёт, выводимый в печатное сообщение вида 8
PLR[1,8]="SELECT * FROM (SELECT '↓ НЕДАВНО ОБНОВЛЁННЫЕ ТАБЛИЦЫ, в которых не по всем показателям/измерениям созда
```

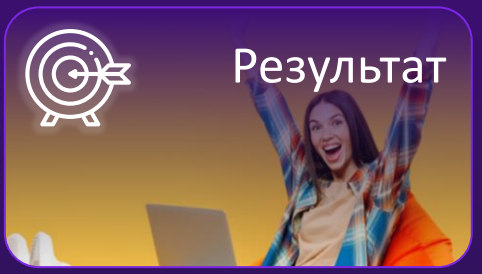
Оповещения об ошибках в документации



Автоматическая работа системы DQ







Безопасный транспорт ЦОДД

Это победа!

10 основных видов проверок создаются и переопределяются сами на основе актуальной документации для Hive и Vertica

Можно добавлять произвольные проверки и гибко управлять параметрами автоматических проверок

Проверки сами начинают выполняться автоматически после выката в прод дага обновления таблицы

Можно пересчитать историю результатов проверок при изменении их логики

Оповещения по непройденным проверкам сами приходят адресно всем заинтересованным в телефон сразу же, как возникает проблема

Автоматически выстраиваются зависимости между проверками витрин и источников данных

Визуализирована актуальная история результатов запуска

Документация на ХД всегда в актуальном состоянии и контролируется автоматически

Инновационный центр

ictransport.ru

Безопасный транспорт

Спасибо за внимание!