

Разработка хранилища данных для такси «Везу и точка»

Авторы проекта:

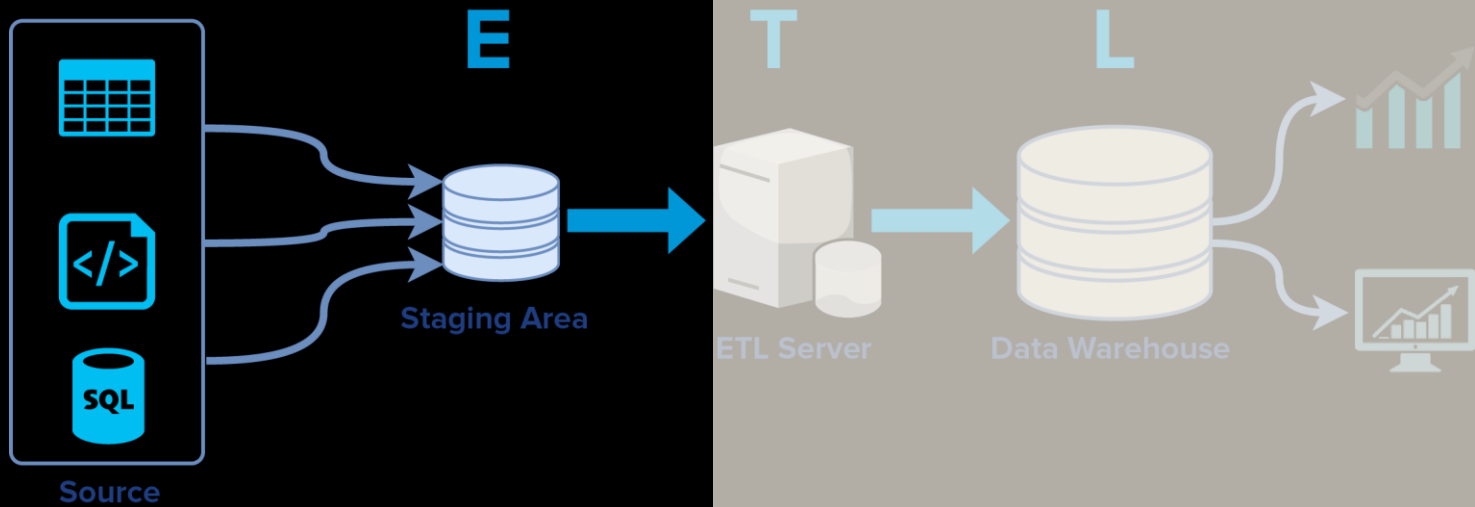
- Камилла Хуранова
- Павел Вервейн
- Тимур Шагимуратов
- Никита Гирш
- Дарья Зайцева



https://github.com/daria-z7/taxi_etl_process

***Инструкция по запуску в README.md**

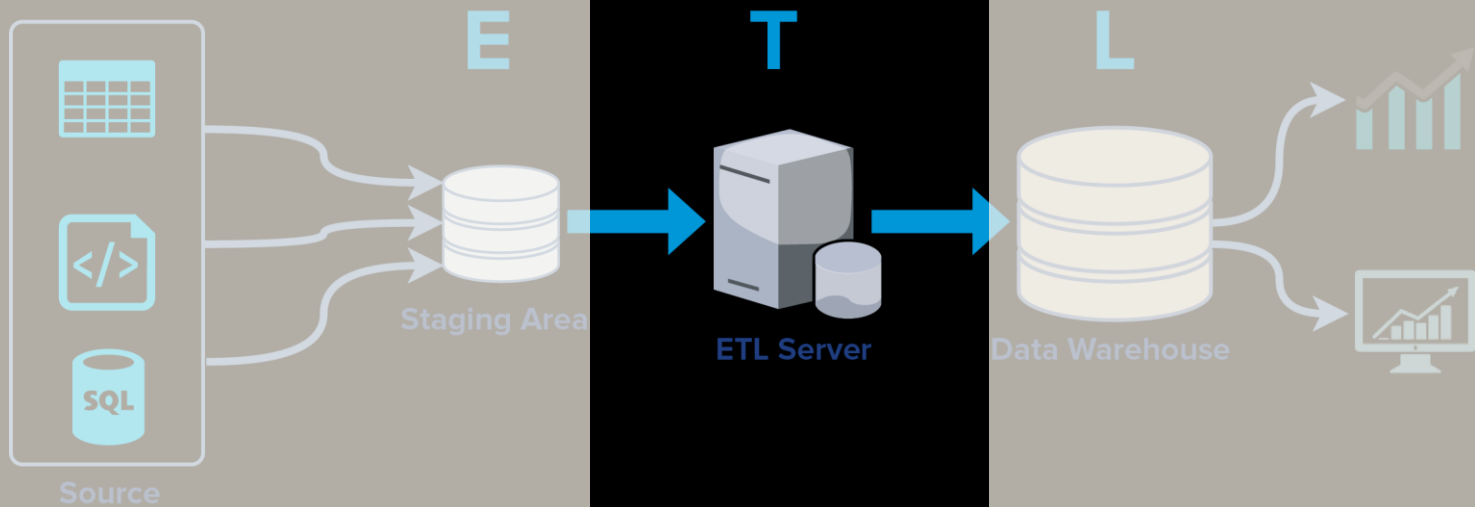
Первый этап



Имеющиеся данные:

- 4 таблицы (main.rides, main.movement, main.car_pool, main.drivers)
- Файловое хранилище (waybills, payments)

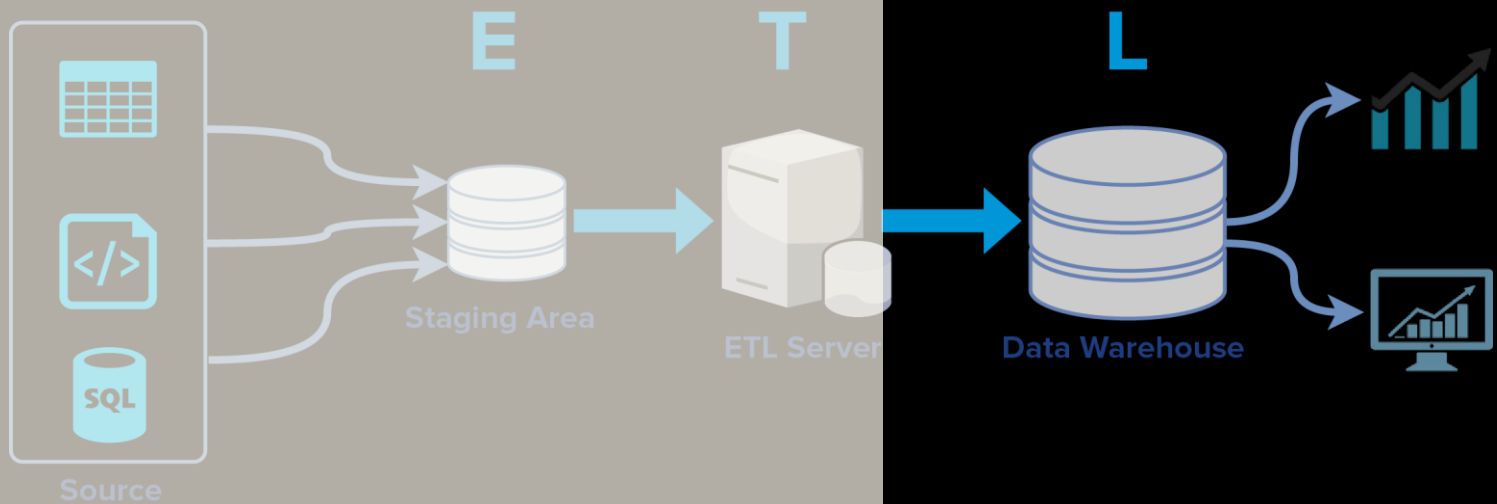
Извлекаем под средством файлов: `load_file_data` и `load_table_data`



Выполненные преобразования:

- Объединение таблиц rides и movement для формирования таблицы fact rides
- Создание новых столбцов для распределения 3-х статусов каждой поездки
- Использование personnel_num из dim_drivers для fact_waybills и fact_rides
- Сопоставление исходных данных с созданными таблицами по проекту:

Фактовые таблицы	Таблицы - измерения
fact_rides	dim_clients
fact_waybills	dim_cars
fact_payments	dim_drivers



Выполненные преобразования:

- Проверка из вспомогательной таблицы `work_load_check` на загрузку данных загружаемого дня
- Во время загрузки формируется «`main.log`» об успешной загрузке в каждую из таблиц
- Пошаговая загрузка данных по таблицам за вчерашний день (сначала таблицы измерений, а потом с фактами, последняя из которых – `fact_rides`). Для определения обновившихся данных опираемся на поля дат (`#update_dt`). Если данные заполнились корректно, то во вспомогательную таблицу `work_load_check` записывается дата загруженного дня
- Безопасность реализована согласно ТЗ

Результат заполнения измерительных таблиц

*<dw_h_saint_petersburg> Script-14 X

```
select *
from dim_clients dc
limit 10;
```

dim_clients 1 X

select * from dim_clients dc limit 10 | Введите SQL выражение чтобы отфильтровать результаты

	asc phone_num	start_dt	asc card_num	asc deleted_flag	end_dt
1	+7 (965) 286-34-10	2022-10-12 12:47:02.000	4627 3019 0269 6662	0	[NULL]
2	+7 (937) 732-54-53	2022-10-12 12:54:01.000	4627 2984 9648 4473	0	[NULL]
3	+7 (935) 927-53-87	2022-10-12 12:54:01.000	4627 3043 2224 6142	0	[NULL]
4	+7 (924) 640-82-46	2022-10-15 20:19:02.000	4478 1856 8505 2858	0	[NULL]
5	+7 (959) 829-75-10	2022-10-12 12:55:02.000	5257 4468 0094 3494	0	[NULL]
6	+7 (967) 767-75-86	2022-10-12 12:57:02.000	4469 5597 5504 3263	0	[NULL]
7	+7 (941) 316-02-50	2022-10-15 20:20:02.000	5254 7782 5578 2677	0	[NULL]
8	+7 (901) 877-22-05	2022-10-15 20:21:01.000	5100 8292 4525 1152	0	[NULL]
9	+7 (967) 678-30-22	2022-10-12 13:05:02.000	5331 5746 5520 3761	0	[NULL]
10	+7 (994) 090-49-23	2022-10-12 13:05:02.000	5445 7332 8555 5262	0	[NULL]

*<dw_h_saint_petersburg> Script-14 X

```
select *
from dim_cars dc
limit 10;
```

dim_cars 1 X

select * from dim_cars dc limit 10 | Введите SQL выражение чтобы отфильтровать результаты

	asc plate_num	start_dt	asc model_name	revision_dt	asc deleted_flag	end_dt
1	K405AP177	2022-10-16 06:04:03.000	Volkswagen Polo	2022-10-19	N	[NULL]
2	C644HX797	2022-10-16 08:11:03.000	Volkswagen Polo	2022-10-19	N	[NULL]
3	K399EM777	2022-10-16 08:24:03.000	Kia Rio	2022-10-19	N	[NULL]
4	O002YE977	2022-10-16 08:40:03.000	Hyundai Solaris	2022-10-19	N	[NULL]
5	C079P0750	2022-10-18 16:25:03.000	Kia Rio	2022-10-21	N	[NULL]
6	X913EP50	2022-10-18 08:03:03.000	Kia Rio	2022-10-21	N	[NULL]
7	C825AC77	2022-10-18 08:13:03.000	Hyundai Solaris	2022-10-21	N	[NULL]
8	O946TX197	2022-10-18 09:06:03.000	Volkswagen Polo	2022-10-21	N	[NULL]
9	M423KT197	2022-10-18 09:07:03.000	Volkswagen Polo	2022-10-21	N	[NULL]
10	M027MH199	2022-10-18 09:15:03.000	Volkswagen Polo	2022-10-21	N	[NULL]

*<dw_h_saint_petersburg> Script-14 X

```
select *
from dim_drivers dd
limit 10;
```

dim_drivers 1 X

select * from dim_drivers dd limit 10 | Введите SQL выражение чтобы отфильтровать результаты

	123 personnel_num	start_dt	asc last_name	asc first_name	asc middle_name	birth_dt	asc card_num	asc driver_license_num	driver_license_dt	asc deleted_flag	end_dt
1	1	2022-10-12 12:47:03.000	Чемиренко	Даниэль	Артемьевич	1968-01-30	4205 4648 9442 7155	13 51 130647	2023-04-12	N	[NULL]
2	2	2022-10-12 12:50:03.000	Головцов	Рослав	Рославович	1976-11-04	7719 1583 7863 4759	38 10 977977	2023-04-12	N	[NULL]
3	3	2022-10-12 12:52:03.000	Сакобов	Алан	Тихонович	1978-08-18	1875 8923 3744 8928	34 48 464847	2023-04-12	N	[NULL]
4	4	2022-10-12 12:54:03.000	Тарлов	Игорь	Ефимович	1993-12-28	2186 7201 8599 1460	42 30 001788	2023-04-12	N	[NULL]
5	5	2022-10-12 12:54:03.000	Свеноков	Богдан	Юсуфович	1981-06-05	9590 3819 7544 6557	31 39 276519	2023-04-12	N	[NULL]
6	6	2022-10-12 12:55:03.000	Попелицкий	Руслан	Германович	1986-02-18	5198 9779 9847 9503	91 95 521602	2023-04-12	N	[NULL]
7	7	2022-10-12 12:55:03.000	Умралиев	Родион	Эрикович	1973-01-09	1486 9214 3856 5510	68 87 611785	2023-04-12	N	[NULL]
8	8	2022-10-12 12:57:03.000	Бондарев	Самир	Германович	1996-05-13	6222 8600 4484 3992	99 00 949942	2023-04-12	N	[NULL]
9	9	2022-10-12 12:57:03.000	Цетнарский	Эмирхан	Ильясович	1993-09-09	5247 4495 9014 2398	22 79 062689	2023-04-12	N	[NULL]
10	10	2022-10-12 12:58:03.000	Даклеев	Тигран	Михайлович	1984-01-08	3503 3829 4377 8481	96 34 555121	2023-04-12	N	[NULL]

Результат заполнения фактовых таблиц

*<dw_h_saint_petersburg> Script-14 X

```
select *
from fact_waybills fw
limit 10;
```

fact_waybills 1 X

select * from fact_waybills fw limit 10 Введите SQL выражение чтобы отфильтровать результаты

	asc waybill_num	123 driver_pers_num	asc car_plate_num	work_start_dt	work_end_dt	issue_dt
1	AA-001		1 T561CO77	2022-10-12 12:00:00.000	2022-10-12 19:00:00.000	2022-10-12
2	AA-002		2 P323BT77	2022-10-12 12:00:00.000	2022-10-12 14:30:00.000	2022-10-12
3	AA-003		3 P071XC77	2022-10-12 12:00:00.000	2022-10-12 18:30:00.000	2022-10-12
4	AA-004		5 P023EH77	2022-10-12 12:00:00.000	2022-10-12 18:30:00.000	2022-10-12
5	AA-005		4 O578PH77	2022-10-12 12:00:00.000	2022-10-12 19:30:00.000	2022-10-12
6	AA-006		7 M506CH77	2022-10-12 12:00:00.000	2022-10-12 19:00:00.000	2022-10-12
7	AA-007		6 A591EV77	2022-10-12 12:00:00.000	2022-10-12 15:00:00.000	2022-10-12
8	AA-008		9 K389TV77	2022-10-12 12:00:00.000	2022-10-12 15:30:00.000	2022-10-12
9	AA-009		8 M919HP77	2022-10-12 12:00:00.000	2022-10-12 18:30:00.000	2022-10-12
10	AA-010		10 P545HA77	2022-10-12 12:00:00.000	2022-10-12 18:00:00.000	2022-10-12

*<dw_h_saint_petersburg> Script-14 X

```
select *
from fact_payments fp
limit 10;
```

fact_payments 1 X

select * from fact_payments fp limit 10 Введите SQL выражение чтобы отфильтровать результаты

	123 transaction_id	asc card_num	123 transaction_amt	transaction_dt
1		1 4438885142042652	124,35	2022-12-10 15:32:03.000
2		2 5257449309434528	259,8	2022-12-10 15:34:02.000
3		3 5254771632413264	37,05	2022-12-10 15:37:03.000
4		4 4058701801921954	314,55	2022-12-10 15:49:03.000
5		5 5257442864273548	386,55	2022-12-10 15:49:03.000
6		6 5545625404698149	252,75	2022-12-10 15:51:03.000
7		7 4104621272619784	310,35	2022-12-10 15:16:03.000
8		8 437773377191027	409,5	2022-12-10 15:21:03.000
9		9 5264834951655549	236,1	2022-12-10 15:21:03.000
10		10 4406665055532628	302,4	2022-12-10 15:24:03.000

*<dw_h_saint_petersburg> Script-14 X

```
select *
from work_load_check wlc
limit 10;
```

work_load_check 1 X

select * from work_load_check wlc limit 10

	123 id_load	date_load
1	9	2022-10-13

*<dw_h_saint_petersburg> Script-14 X

```
select *
from fact_rides fr
limit 10;
```

fact_rides 1 X

select * from fact_rides fr limit 10 Введите SQL выражение чтобы отфильтровать результаты

	123 ride_id	asc point_from_txt	asc point_to_txt	123 distance_val	123 price_amnt	asc client_phone_num	123 driver_pers_num	asc car_plate_num	ride_arrival_dt	ride_start_dt	ride_end_dt
1	7	7-я улица Лазенки,	улица Удальцова,	22,3	334,5	+7 (959) 829-75-10	6 A591EV77		2022-10-12 13:02:03.000	[NULL]	2022-10-12 13:03:02.000
2	3	Большая Академиче	Алтуфьевское шос	10,49	157,35	+7 (941) 239-10-32	3 P071XC77		2022-10-12 12:55:04.000	2022-10-12 12:56:03.000	2022-10-12 13:04:03.000
3	10	Буженинова улица,	Большой Афанась	22,29	334,35	+7 (999) 574-71-59	8 M919HP77		2022-10-12 13:03:02.000	[NULL]	2022-10-12 13:05:03.000
4	15	Мосфильмовская ул	2-й Железнодорож	25,09	376,35	+7 (994) 090-49-23	13 P030PC99		2022-10-12 13:09:03.000	[NULL]	2022-10-12 13:09:03.000
5	12	Долгоруковская ули	Малая Семёновск	23,29	349,35	+7 (942) 465-22-39	8 M919HP77		2022-10-12 13:07:03.000	[NULL]	2022-10-12 13:10:03.000
6	16	проспект Мира, 71 с	улица Шаболовка,	20,76	311,4	+7 (959) 829-75-10	12 E329PP150		2022-10-12 13:09:03.000	[NULL]	2022-10-12 13:12:03.000
7	2	Волгоградский прос	Рабочая улица, 84	14,22	213,3	+7 (974) 683-22-78	2 P323BT77		2022-10-12 12:55:04.000	2022-10-12 12:58:03.000	2022-10-12 13:13:03.000
8	9	Большая Черёмушк	улица Красина, 17	7,83	117,45	+7 (952) 618-15-96	6 A591EV77		2022-10-12 13:04:03.000	2022-10-12 13:05:03.000	2022-10-12 13:13:03.000
9	1	Валовая улица, 11-1	Луговая улица, 6А	18,81	282,15	+7 (965) 286-34-10	1 T561CO77		2022-10-12 12:50:03.000	2022-10-12 12:56:03.000	2022-10-12 13:14:03.000
10	20	проспект Мира, 105	Рабочая улица, 24	3,07	46,05	+7 (956) 671-24-53	14 C214TP777		2022-10-12 13:13:03.000	[NULL]	2022-10-12 13:15:03.000

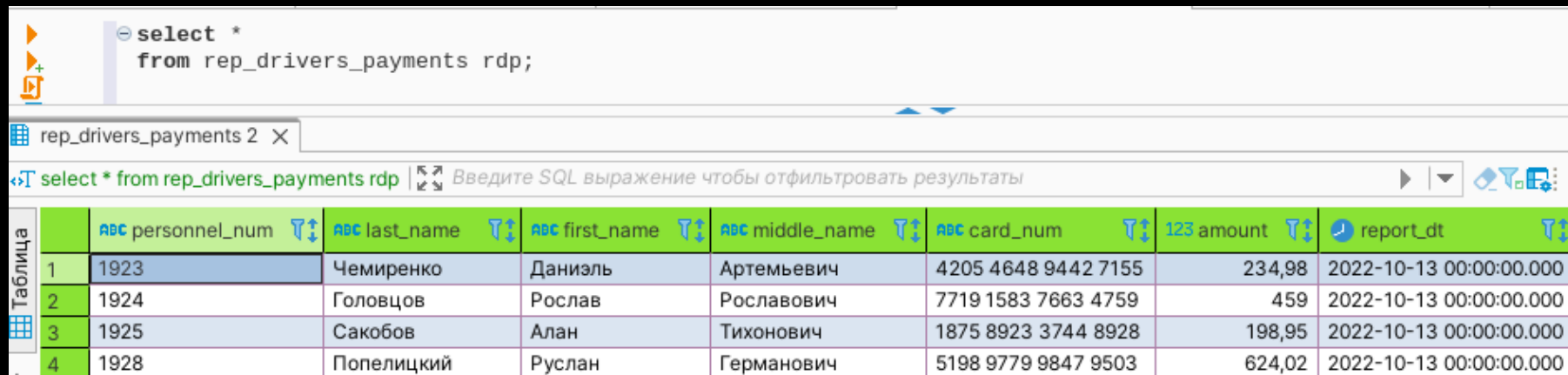
Результаты по первому этапу:

- Реализовали хранилище данных для такси
- Выполнили условия ТЗ
- Данные имеют удобный формат для составления будущих отчетов и дальнейшей аналитики



Второй этап

Таблица (rep_drivers_payments)



The screenshot shows a SQL client interface. At the top, a query editor contains the SQL statement: `select * from rep_drivers_payments rdp;`. Below the editor, a tab labeled "rep_drivers_payments 2" is active. A filter bar shows the same query: `select * from rep_drivers_payments rdp`. The results are displayed in a table with 8 columns: personnel_num, last_name, first_name, middle_name, card_num, amount, and report_dt. The table contains 4 rows of data. The interface includes standard SQL tool icons like execute, refresh, and save.

	personnel_num	last_name	first_name	middle_name	card_num	amount	report_dt
1	1923	Чемиренко	Даниэль	Артемович	4205 4648 9442 7155	234,98	2022-10-13 00:00:00.000
2	1924	Головцов	Рослав	Рославович	7719 1583 7663 4759	459	2022-10-13 00:00:00.000
3	1925	Сакобов	Алан	Тихонович	1875 8923 3744 8928	198,95	2022-10-13 00:00:00.000
4	1928	Попелицкий	Руслан	Германович	5198 9779 9847 9503	624,02	2022-10-13 00:00:00.000

- Условия были выполнены относительно T3
- Report dt – это дата + 1 по имеющимся данным

Таблица (rep_drivers_violations)

select *
from rep_drivers_violations rdv;

rep_drivers_violations 2 X

select * from rep_drivers_violations rdv Введите SQL выражение чтобы отфильтровать резу

	personnel_num	ride	speed	violations_cnt	
1	1923	897	88,5	0	
2	1923	3856	88,2	1	
3	1923	10612	85,0285714286	2	
4	1923	14288	88,1294117647	3	
5	1923	14472	87,1714285714	4	
6	1923	24751	102,8	5	
7	1923	24904	85,6	6	
8	1923	27295	94,3285714286	7	
9	1923	27417	87,675	8	
10	1923	27512	93,5076923077	9	
11	1924	6605	85,7142857143	0	
12	1925	1071	86,24	0	
13	1925	1405	87,3	1	
14	1925	2832	91,48	2	

- Средняя скорость была рассчитана $\text{distance} / (\text{end_dt} - \text{start_dt})$ в расчёте на км
- Для счетчика была использована оконная функция `row_number - 1`

Таблица (rep_drivers_overtime)

```
select *  
from rep_drivers_overtime rdo;
```

rep_drivers_overtime 1 X

select * from rep_drivers_overtime rdo | Введите SQL выражение чтобы отфильтровать

	ABC personnel_num	start_waybill	sum_work_time
1	1923	2022-10-17 07:00:00.000	07:30:00
2	1923	2022-10-25 02:00:00.000	08:00:00
3	1925	2022-10-12 21:00:00.000	08:00:00
4	1926	2022-10-13 04:00:00.000	08:00:00
5	1927	2022-10-20 05:00:00.000	07:30:00
6	1928	2022-10-12 15:00:00.000	11:00:00
7	1932	2022-10-13 06:00:00.000	07:30:00

- Была учтена ошибка, что путевые листы могут накладываться друг на друга. Был использован min max в разрезе промежутка. Пример визуализации внизу, где красные стрелки – суммарные наработки.

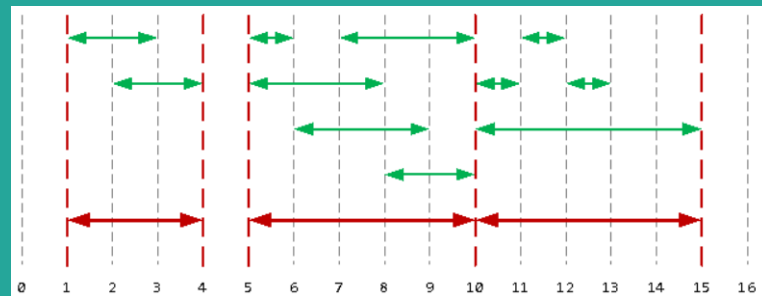
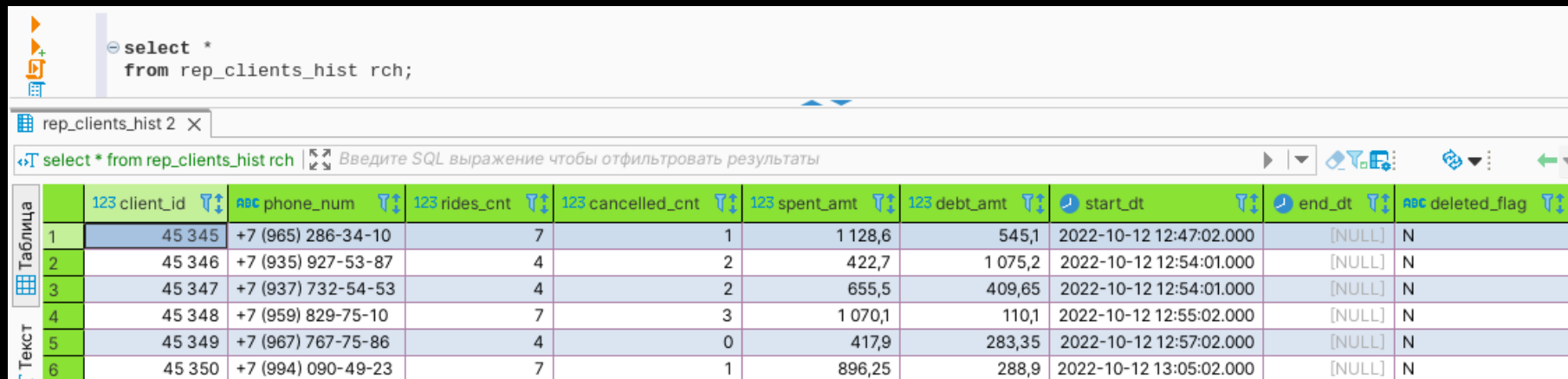


Таблица (rep_clients_hist)



The screenshot shows a SQL client window with a query editor at the top and a results grid below. The query is `select * from rep_clients_hist rch;`. The results grid has 10 columns: `client_id`, `phone_num`, `rides_cnt`, `cancelled_cnt`, `spent_amt`, `debt_amt`, `start_dt`, `end_dt`, and `deleted_flag`. There are 6 rows of data. The interface includes a toolbar with various icons for filtering and sorting, and a search bar with the placeholder text "Введите SQL выражение чтобы отфильтровать результаты".

	123 client_id	abc phone_num	123 rides_cnt	123 cancelled_cnt	123 spent_amt	123 debt_amt	start_dt	end_dt	abc deleted_flag
1	45 345	+7 (965) 286-34-10	7	1	1 128,6	545,1	2022-10-12 12:47:02.000	[NULL]	N
2	45 346	+7 (935) 927-53-87	4	2	422,7	1 075,2	2022-10-12 12:54:01.000	[NULL]	N
3	45 347	+7 (937) 732-54-53	4	2	655,5	409,65	2022-10-12 12:54:01.000	[NULL]	N
4	45 348	+7 (959) 829-75-10	7	3	1 070,1	110,1	2022-10-12 12:55:02.000	[NULL]	N
5	45 349	+7 (967) 767-75-86	4	0	417,9	283,35	2022-10-12 12:57:02.000	[NULL]	N
6	45 350	+7 (994) 090-49-23	7	1	896,25	288,9	2022-10-12 13:05:02.000	[NULL]	N

- Для составления связей соединили таблицы: `dim_clients`, `fact_rides`, `fact_payments`
- Если оплата не прошла, то мы добавляли `price_amt` в долг клиента
- Учили независимость используемой карты для начисления долга и потраченной суммы на поездки

Результаты по второму этапу:

- Корректировка загрузки данных :
 - Отладили правильность заполнения водителей в fact_rides
 - Исправили неточность версий в dim_clients
 - Исправили время закрытия предыдущей версии от времени начала следующей путем вычитания 1 сек
 - Заменяли в фильтрации like на =
 - Исправили функцию get_personnel_num для водителя, где фильтрация не только по номеру машины, но и по времени привязки машины к водителю
 - Выполнили условия ТЗ по 4 таблицам и учли возможные ошибки
-