Версионное хранение данных

Типы версионностей

Тригубова Оксана

Архитектор данных

Oksana.Trigubova@X5.ru

О чем поговорим?

1. Что такое техническая история?
2. Что такое бизнес-история? Рассмотрим на примере ассортимента в магазине
3. Двойная история
4. В каких объектах может быть историчность?
5. PIT – зачем он нужен?
6. Примеры работы с историей

Виды историчности

По типу

1. Техническая
2. Бизнес
3. Двойная

По способу хранения

1. SCD0 - значения атрибутов не будут меняться
2. SCD1 – значения атрибутов полностью заменяются на новые
3. SCD2 - добавление новой строки с сохранением предыдущей и сохранение в дополнительных столбцах признаков актуальности. Такой подход позволяет сохранить историчность.
4. Срезы

Техническая история

1. Запись с источника –

plu_id	weight_unit	brutto	hwd_unit	height	width	depth
4032191	кг	0.469	см	1,68	0,63	0,63

2. Загружаем в

plu_id	weight_uom_code	plu_brutto_weight_amt	hwd_uom_code	plu_height_amt	plu_width_amt	plu_depth_amt	valid_from_dttm	actual_flg
4032191	KG	0.469					1960-01-01 00:00:00	0
4032191	KG	0.469	см	1.68	0.63	0.63	2020-02-14 00:00:00	1

3. Запись изменилась на источнике –

plu_id	weight_unit	brutto	hwd_unit	height	width	depth
4032191	кг	0,469	см	16,8	6,3	6,3

4. Делаем версионность в

plu_id	weight_uom_code	plu_brutto_weight_amt	hwd_uom_code	plu_height_amt	plu_width_amt	plu_depth_amt	valid_from_dttm	actual_flg
4032191	KG	0.469					1960-01-01 00:00:00	0
4032191	KG	0.469	см	1.68	0.63	0.63	2020-02-14 00:00:00	0
4032191	KG	0.469	см	16.8	6.3	6.3	2020-02-18 00:00:00	1

Техническая история

- это хронология изменения атрибутов какого-либо объекта
- происходит вследствие обновления записи на источнике или расчета нового значения

Непрерывность:

от даты "сотворения мира" - '01-01-1960' ('01-01-1960 00:00:00')

до даты "апокалипсиса" - '01-01-5999' ('01-01-5999 00:00:00')

Техническая история

Разбор инцидентов

Восстановление отчетности

Нужна пользователям, когда отражает бизнес-процесс

Бизнес - история

ассортимента в магазине

- Источник
  ERP:

datab	datbi	asort	name1	locnr
14.06.2016	31.07.2016	CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154
17.01.2018	11.09.2018	CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154

- После загрузки в EDW -

business_from_dttm	business_to_dttm	hashdiff_key	assortment_erp_id	assortment_nm	store_id
01.01.1960 0:00	13.06.2016 23:59		CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154
14.06.2016 0:00	**31.07.2016 23:59**	d41d8cd98f00b	CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154
01.08.2016 0:00	16.01.2018 23:59		CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154
17.01.2018 0:00	**11.09.2018 23:59**	d41d8cd98f00b	CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154
12.09.2018 0:00	01.01.5999 0:00		CHLIS14525	СМ ЛИП Сервировочная посуда_25	2154

Бизнес - история

рассчитываемая

- Даты действия товара в ассортименте –

business_from_dttm	business_to_dttm	assortment_id	assortment_nm	plu_id	assortment_plu_type_dk	plu_nm
31.01.2017 0:00	01.01.5999 0:00	L310	5994-Пятерочка	3273501	5	Пиво PILSNER URQUELL
29.07.2013 0:00	01.01.5999 0:00	L310	5994-Пятерочка	3273501	7	Пиво PILSNER URQUELL

- Даты действия локального ассортимента – GRP_RV.S_ASSORTMENT_WRS1_ERP

assortment_start_dttm	assortment_end_dttm	assortment_id	assortment_nm
18.08.2015 0:00	06.02.2017 23:59	L310	5994-Пятерочка

- Результат расчета в Bridge -

business_from_dttm	business_to_dttm	assortment_id	assortment_nm	plu_id	assortment_plu_type_dk	plu_nm
31.01.2017 0:00	06.02.2017 23:59	L310	5994-Пятерочка	3273501	5	Пиво PILSNER URQUELL
18.08.2015 0:00	06.02.2017 23:59	L310	5994-Пятерочка	3273501	7	Пиво PILSNER URQUELL

Бизнес - история

– хронология изменения атрибутов, отражающие время их действия, описывающих бизнес-процесс.

Непрерывность:

от даты "сотворения мира" - '01-01-1960' ('01-01-1960 00:00:00')

до даты "апокалипсиса" - '01-01-5999' ('01-01-5999 00:00:00')

Двойная история

Полный ключ: **valid_from_dttm, link_key, business_from_dttm**

Бизнес-ключ: **link_key, business_from_dttm**

Товар в ассортименте GRP_RV.M_ASSORTMENT_X_PLU_WLK1_ERP

actual_flg	valid_from_dttm	assortment_nm	plu_nm	business_from_dttm	business_to_dttm
	01.01.1960 0:00	8540-Пятерочка	ЧЕР.КАРТА Кофе GOLD раст.субл.75г	29.06.2020 0:00	
0	30.06.2020 0:00	8540-Пятерочка	ЧЕР.КАРТА Кофе GOLD раст.субл.75г	29.06.2020 0:00	01.01.5999 0:00
1	17.08.2020 0:00	8540-Пятерочка	ЧЕР.КАРТА Кофе GOLD раст.субл.75г	29.06.2020 0:00	15.08.2020 23:59

Ассортимент в магазине GRP_RV.S_ASSORTMENT_X_STORE_WRSZ_ERP

actual_flg	valid_from_dttm	link_key	assortment_nm	store_nm	business_from_dttm	business_to_dttm	hashdiff_key
1	02.07.2020 0:00	**00034b79f5d359a**	ДСК ВОР Конф.Фас	663-Пятерочка	**01.01.1960 0:00**	02.06.2019 23:59	
0	02.07.2020 0:00	**00034b79f5d359a**	ДСК ВОР Конф.Фас	663-Пятерочка	**03.06.2019 0:00**	01.01.5999 0:00	d41d8cd98f00b
1	07.07.2020 0:00	**00034b79f5d359a**	ДСК ВОР Конф.Фас	663-Пятерочка	**03.06.2019 0:00**	20.07.2020 23:59	d41d8cd98f00b
1	07.07.2020 0:00	**00034b79f5d359a**	ДСК ВОР Конф.Фас	663-Пятерочка	**21.07.2020 0:00**	01.01.5999 0:00	

Связь существует, но не действует

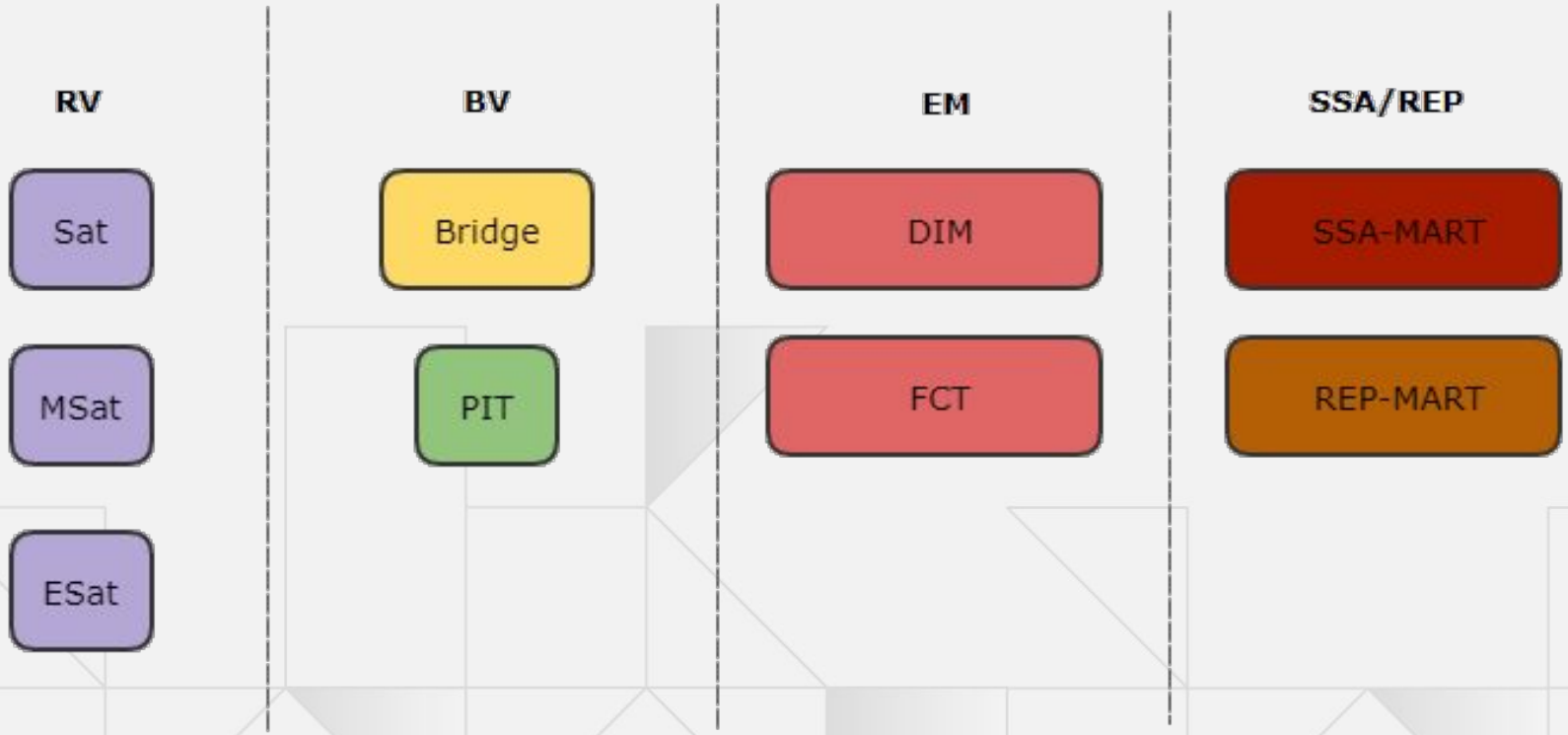
Какая историчность возможна?

Данные	Имеют бизнес-хронологию	Не имеют бизнес-хронологию
Обновляются	Двойная	Техническая
Не обновляются	Бизнес история	Без истории

Хранение по способу извлечения

	Полная история	История снимками
Формат хранения в источнике	В источнике история представлена в виде полного лога всех изменений загружаемой таблицы.	Только актуальный срез
Способ извлечения из источника	Использование механизма CDC для репликации источников в контур хранилища с сохранением полного лога изменений.	1) Инкрементальная загрузка с вычислением инкремента по полю со временем изменения записи 2) Загрузка таблицы-источника целиком с последующим полным сравнением с уже загруженными данными
Формат хранения в EDW	Временые интервалы актуальности версий определены с точностью до секунды.	Временные интервалы актуальности версий округляются до частоты загрузки (для загрузок 1/сутки округление до даты).

В каких объектах может быть историчность?



Point-in-Time (PIT)

– предназначен для построения сводной истории изменения атрибутов сущности.

Критерии создания

- Пересечение истории, из нескольких сателлитов/бриджей сущности
- Расчет интервалов актуальности записей (valid_from_dttm → valid_from_dttm + valid_to_dttm)

Point-in-Time (PIT)

Структура таблицы

Key	Поле	Описание
	dataflow_id	ID процесса, обработавшего запись
	dataflow_dttm	Дата и время обработки записи процессом
PK	<*entity_name*>_rk	Постоянный ключ сущности, для которой строится история
PK	valid_from_dttm	Дата и время начала интервала актуальности записи
	valid_to_dttm	Дата и время окончания интервала актуальности записи
	<*satellite_name_1*>_vf_dttm	Значение поля valid_from_dttm из записи 1-го сателлита , соответствующей интервалу актуальности PIT-таблицы.
	...	
	<*satellite_name_m*>_vf_dttm	Значение поля valid_from_dttm из записи N-го сателлита , соответствующей интервалу актуальности PIT-таблицы.

Point-in-Time (PIT

Структура таблицы для двойной версионности

Key	Поле	Описание
	dataflow_id	ID процесса, обработавшего запись
	dataflow_dttm	Дата и время обработки записи процессом
PK	*<entity_name>*_rk	Постоянный ключ сущности, для которой строится история
PK	valid_from_dttm	Дата и время начала интервала актуальности записи
	valid_to_dttm	Дата и время окончания интервала актуальности записи
PK	business_from_dttm	Дата/время открытия бизнес версии **Важно!** Наличие поля – опционально.
	business_to_dttm	Дата/время закрытия бизнес версии **Важно!** Наличие поля – опционально.
	*<satellite_name_1>*_vf_dttm	Значение поля valid_from_dttm из записи 1-го сателлита , соответствующей интервалу актуальности PIT-таблицы.
	*<satellite_name_1>*_bf_dttm	Значение поля business_from_dttm из записи 1-го сателлита , соответствующей интервалу актуальности PIT-таблицы. **Важно!** Наличие поля – опционально.
	...	
	*<satellite_name_m>*_vf_dttm	Значение поля valid_from_dttm из записи M-го сателлита , соответствующей интервалу актуальности PIT-таблицы.
	*<satellite_name_m>*_bf_dttm	Значение поля business_from_dttm из записи M-го сателлита , соответствующей интервалу актуальности PIT-таблицы. **Важно!** Наличие поля – опционально.

Point-in-Time (PIT)

store_rk	valid_from_dttm	store_nm	dist_chan_code	division_code	macroregion_code
a9deb025d9c	01.01.1960 0:00				
a9deb025d9c	02.10.2020 0:00	20590-Пятерочка	D1		
a9deb025d9c	03.10.2020 0:00	20590-Пятерочка	D1	1400	MRDVO

SAL

store_rk	valid_from_dttm	dist_org_code
bf5bbfe272c1b9a	01.01.1960 0:00	
bf5bbfe272c1b9a	02.10.2020 0:00	NG21

Как формируется история в PIT-таблице

			SAT1					SAT2	
store_rk	valid_from_dttm	valid_to_dttm	s_1_mdm_vf_dttm	store_nm	dist_chan_code	division_code	macroregion_code	s_2_erp_vf_dttm	dist_org_code
a9deb025d9cf	01.01.1960 0:00	01.10.2020 23:59	01.01.1960 0:00					01.01.1960 0:00	
a9deb025d9cf	02.10.2020 0:00	02.10.2020 23:59	02.10.2020 0:00	20590-Пятерочка	D1			02.10.2020 0:00	NG21
a9deb025d9cf	03.10.2020 0:00	01.01.5999 0:00	03.10.2020 0:00	20590-Пятерочка	D1	1400	MRDVO	02.10.2020 0:00	NG21

PIT

store_rk	valid_from_dttm	valid_to_dttm	s_store_werks_mdm_vf_dttm	s_store_t001w_erp_vf_dttm
a9deb025d9cf50ec6a	01.01.1960 0:00	01.10.2020 23:59	01.01.1960 0:00	01.01.1960 0:00
a9deb025d9cf50ec6a	02.10.2020 0:00	02.10.2020 23:59	02.10.2020 0:00	02.10.2020 0:00
a9deb025d9cf50ec6a	03.10.2020 0:00	01.01.5999 0:00	03.10.2020 0:00	02.10.2020 0:00

Point-in-Time (PIT)

assortment_rk	valid_from_dttm	assortment_erp_id	assortment_start_dttm	assortment_end_dttm
e3dc985bffbdd8972	01.01.1960 0:00			
e3dc985bffbdd8972	03.07.2020 0:00	L626	23.09.2015 0:00	01.01.5999 0:00
e3dc985bffbdd8972	17.07.2020 0:00	L626	23.09.2015 0:00	08.06.2020 23:59

SAL

assortment_rk	valid_from_dttm	assortment_nm
e3dc985bffbdd8972	01.01.1960 0:00	
e3dc985bffbdd8972	21.02.2020 0:00	5908-Пятерочка

Как формируется история в PIT-таблице

			SAT1				SAT2	
assortment_rk	valid_from_dttm	valid_to_dttm	s_assort_wrs1_erp_vf_dttm	assort_erp_id	assort_start_dttm	assort_end_dttm	s_assort_wrst_erp	assortment_nm
e3dc985bffbdd897	01.01.1960 0:00	20.02.2020 23:59	01.01.1960 0:00				01.01.1960 0:00	
e3dc985bffbdd897	21.02.2020 0:00	02.07.2020 23:59	01.01.1960 0:00				21.02.2020 0:00	5908-Пятерочка
e3dc985bffbdd897	03.07.2020 0:00	16.07.2020 23:59	03.07.2020 0:00	L626	23.09.2015 0:00	01.01.5999 0:00	21.02.2020 0:00	5908-Пятерочка
e3dc985bffbdd897	17.07.2020 0:00	01.01.5999 0:00	17.07.2020 0:00	L626	23.09.2015 0:00	08.06.2020 23:59	21.02.2020 0:00	5908-Пятерочка

PIT

assortment_rk	valid_from_dttm	valid_to_dttm	s_assortment_wrs1_erp_vf_dttm	s_assortment_wrst_erp_vf_dttm
e3dc985bffbdd897	01.01.1960 0:00	20.02.2020 23:59	01.01.1960 0:00	01.01.1960 0:00
e3dc985bffbdd897	03.07.2020 0:00	16.07.2020 23:59	03.07.2020 0:00	21.02.2020 0:00
e3dc985bffbdd897	21.02.2020 0:00	02.07.2020 23:59	01.01.1960 0:00	21.02.2020 0:00
e3dc985bffbdd897	17.07.2020 0:00	01.01.5999 0:00	17.07.2020 0:00	21.02.2020 0:00

Примеры работы с историей

Алгоритм перенарезки истории

Дано товар в магазине в нескольких типах ассортимента
GRP_BV.B_ASSORTMENT_X_PLU_X_STORE

plu_id	plu_nm	store_id	store_nm	assortment_nm	assortment_type_dk	assortment_type_desc	business_from_dttm	business_to_dttm
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	ДСК ТУЙ Водка_05	GNRL	Общий ассортимент	27.10.2016 0:00	27.10.2016 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	Алкоголь лицензируемый	ISKL	Исключения товарных групп	29.10.2016 0:00	21.02.2017 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	ДСК ТУЙ Водка+Коньяк_05	GNRL	Общий ассортимент	21.11.2016 0:00	30.04.2019 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	ДСК ТУЙ Водка+Коньяк_05	GNRL	Общий ассортимент	01.05.2019 0:00	01.01.5999 0:00
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	8653-Пятерочка	PROM	Рекламный ассортимент	02.06.2020 0:00	30.06.2020 23:59

Что хотим
получить:

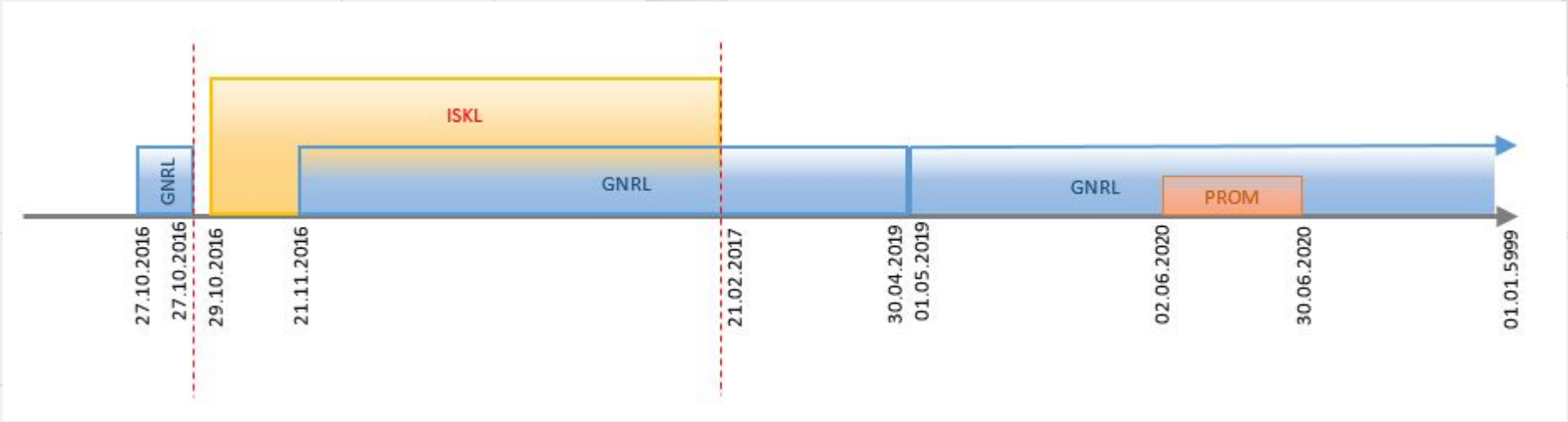
plu_id	plu_nm	store_id	store_nm	sale_plan_flg	business_from_dttm	business_to_dttm
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	1	27.10.2016 0:00	27.10.2016 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	0	28.10.2016 0:00	21.02.2017 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	1	22.02.2017 0:00	01.01.5999 0:00

Алгоритм перенарезки истории

1. ORIGINAL_INTERVALS: Разметили типы ассортимента – определили

plu_id	plu_nm	store_id	store_nm	assortment_nm	sale_plan_flg	business_from_dttm	business_to_dttm
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	ДСК ТУЙ Водка_05	1	27.10.2016 0:00	27.10.2016 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	Алкоголь лицензируемый	0	29.10.2016 0:00	21.02.2017 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	ДСК ТУЙ Водка+Коньяк_05	1	21.11.2016 0:00	30.04.2019 23:59
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	ДСК ТУЙ Водка+Коньяк_05	1	01.05.2019 0:00	01.01.5999 0:00
3166498	Водка КУРАЙ 40% д	J372	8653-Пятерочка	8653-Пятерочка	1	02.06.2020 0:00	30.06.2020 23:59

* Для разметки должна быть использована TUNE - таблица

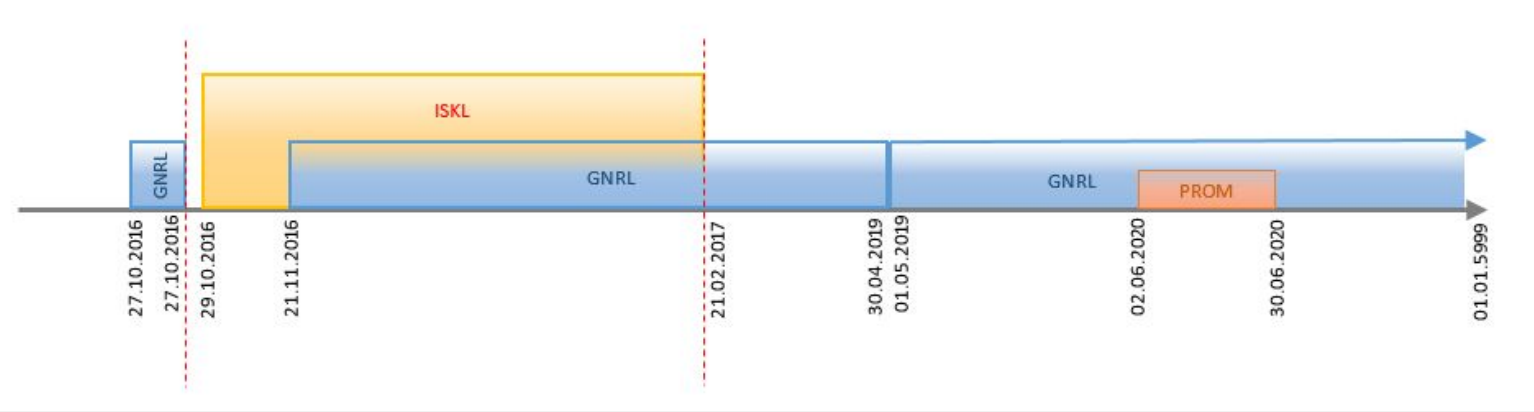


Алгоритм перенарезки истории

2. CALENDAR: Разворачиваем исходные интервалы в календарь

```
SELECT PLU_RK
     , STORE_RK
     , BUSINESS_FROM_DTTM CAL_DTTM
FROM ORIGINAL_INTERVALS
UNION
SELECT PLU_RK
     , STORE_RK
     , BUSINESS_TO_DTTM + INTERVAL '1 SECOND'
FROM ORIGINAL_INTERVALS
WHERE BUSINESS_TO_DTTM <> '5999-01-01'::TIMESTAMP
```

plu_rk	store_rk	cal_dttm
b69d3960b3	cffdd6f3bc	27.10.2016 0:00
b69d3960b3	cffdd6f3bc	28.10.2016 0:00
b69d3960b3	cffdd6f3bc	29.10.2016 0:00
b69d3960b3	cffdd6f3bc	21.11.2016 0:00
b69d3960b3	cffdd6f3bc	22.02.2017 0:00
b69d3960b3	cffdd6f3bc	01.05.2019 0:00
b69d3960b3	cffdd6f3bc	02.06.2020 0:00
b69d3960b3	cffdd6f3bc	01.07.2020 0:00



Алгоритм перенарезки истории

3. NEW_INTERVALS_WITH_SALE_PLAN_FLG: Расставляем каждому дню календаря флаги с учетом приоритета

```
SELECT C.PLU_RK
     , C.STORE_RK
     , COALESCE(MIN(OI.SALE_PLAN_FLG),0) AS SALE_PLAN_FLG
     , C.CAL_DTTM
FROM CALENDAR C
LEFT JOIN ORIGINAL_INTERVALS OI
  ON C.PLU_RK = OI.PLU_RK
 AND C.STORE_RK = OI.STORE_RK
 AND C.CAL_DTTM >= OI.BUSINESS_FROM_DTTM
 AND C.CAL_DTTM <= OI.BUSINESS_TO_DTTM
GROUP BY C.PLU_RK
       , C.STORE_RK
       , C.CAL_DTTM
```

plu_rk	store_rk	sale_plan_flg	cal_dttm
b69d3960b3	cffdd6f3bc	1	27.10.2016 0:00
b69d3960b3	cffdd6f3bc	0	28.10.2016 0:00
b69d3960b3	cffdd6f3bc	0	29.10.2016 0:00
b69d3960b3	cffdd6f3bc	0	21.11.2016 0:00
b69d3960b3	cffdd6f3bc	1	22.02.2017 0:00
b69d3960b3	cffdd6f3bc	1	01.05.2019 0:00
b69d3960b3	cffdd6f3bc	1	02.06.2020 0:00
b69d3960b3	cffdd6f3bc	1	01.07.2020 0:00

Алгоритм перенарезки истории

4. LG_SALE_PLAN_FLG Присваиваем предыдущее состояние флага

```
SELECT NI2.PLU_RK
     , NI2.STORE_RK
     , NI2.SALE_PLAN_FLG
     , NI2.CAL_DTTM
     , LAG(SALE_PLAN_FLG, 1, -1)
       OVER (PARTITION BY NI2.PLU_RK, NI2.STORE_RK ORDER BY NI2.CAL_DTTM) LG_SALE_PLAN_FLG
FROM NEW_INTERVALS_WITH_ASSORTMENT_TYPE_FLG NI2
```

plu_rk	store_rk	sale_plan_flg	cal_dttm	lg_sale_plan_flg
b69d3960b3	cffdd6f3bc	1	27.10.2016 0:00	-1
b69d3960b3	cffdd6f3bc	0	28.10.2016 0:00	1
b69d3960b3	cffdd6f3bc	0	29.10.2016 0:00	0
b69d3960b3	cffdd6f3bc	0	21.11.2016 0:00	0
b69d3960b3	cffdd6f3bc	1	22.02.2017 0:00	0
b69d3960b3	cffdd6f3bc	1	01.05.2019 0:00	1
b69d3960b3	cffdd6f3bc	1	02.06.2020 0:00	1
b69d3960b3	cffdd6f3bc	1	01.07.2020 0:00	1

Алгоритм перенарезки истории

5. Итог. Нарезаем историю. Убираем строки, у которых состояние флага не изменилось

```
SELECT PLU_RK
     , STORE_RK
     , SALE_PLAN_FLG
     , CAL_DTTM AS BUSINESS_FROM_DTTM
     , LEAD(CAL_DTTM - INTERVAL '1 SECOND', 1, '5999-01-01'::TIMESTAMP)
       OVER (PARTITION BY PLU_RK, STORE_RK ORDER BY CAL_DTTM) BUSINESS_TO_DTTM
 FROM LG_SALE_PLAN_FLG
WHERE LG_SALE_PLAN_FLG != SALE_PLAN_FLG
```

plu_rk	store_rk	sale_plan_flg	business_from_dttm	business_to_dttm
b69d3960b36c5	cffdd6f3bcdd8997	1	27.10.2016 0:00	27.10.2016 23:59
b69d3960b36c5	cffdd6f3bcdd8997	0	28.10.2016 0:00	21.02.2017 23:59
b69d3960b36c5	cffdd6f3bcdd8997	1	22.02.2017 0:00	01.01.5999 0:00

Q&A

Версионное хранение данных

Тригубова Оксана

Архитектор данных

Oksana.Trigubova@X5.ru